Ph.D. Dissertation of Seung Hee Yang

Pronunciation Variation Analysis and CycleGAN-based Feedback Generation for CAPT

CAPT를 위한 발음 변이 분석 및 CycleGAN 기반 피드백 생성

February 2020

Seoul National University
Interdisciplinary Program in Cognitive Science

Seung Hee Yang

Pronunciation Variation Analysis and CycleGAN-based Feedback Generation for CAPT

CAPT를 위한 발음 변이 분석 및 CycleGAN 기반 피드백 생성

지도교수 정 민 화

이 논문을 공학박사 학위논문으로 제출함

2020 년 2 월

서울대학교 대학원

협동과정 인지과학전공

Seung Hee Yang

Seung Hee Yang의 공학박사 학위논문을 인준함

2020 년 1 월

위 원 장 _____ 김 홍 기 _____

부위원장 _____ 정 민 화 _____

위 원 _____ 이 호 영 _____

위 원 _____ 김 선 희 _____

위 원 _____ 김 지 환 _____

Abstract

Despite the growing popularity in learning Korean as a foreign language and the rapid development in language learning applications, the existing computer-assisted pronunciation training (CAPT) systems in Korean do not utilize linguistic characteristics of non-native Korean speech. Pronunciation variations in non-native speech are far more diverse than those observed in native speech, which may pose a difficulty in combining such knowledge in an automatic system. Moreover, most of the existing methods rely on feature extraction results from signal processing, prosodic analysis, and natural language processing techniques. Such methods entail limitations since they necessarily depend on finding the right features for the task and the extraction accuracies.

This thesis presents a new approach for corrective feedback generation in a CAPT system, in which pronunciation variation patterns and linguistic correlates with accentedness are analyzed and combined with a deep neural network approach, so that feature engineering efforts are minimized while maintaining the linguistically important factors for the corrective feedback generation task. Investigations on non-native Korean speech characteristics in contrast with those of native speakers, and their correlation with accentedness judgement show that both segmental and prosodic variations are important factors in a Korean CAPT system.

The present thesis argues that the feedback generation task can be interpreted as a style transfer problem, and proposes to evaluate the idea using

generative adversarial network. A corrective feedback generation model is trained on 65,100 read utterances by 217 non-native speakers of 27 mother tongue backgrounds. The features are automatically learnt in an unsupervised way in an auxiliary classifier CycleGAN setting, in which the generator learns to map a foreign accented speech to native speech distributions. In order to inject linguistic knowledge into the network, an auxiliary classifier is trained so that the feedback also identifies the linguistic error types that were defined in the first half of the thesis. The proposed approach generates a corrected version the speech using the learner's own voice, outperforming the conventional Pitch-Synchronous Overlap-and-Add method.

Keyword : Computer-Assisted Pronunciation Training (CAPT), Linguistic Analysis of Non-native Korean, Corrective Feedback Generation for Language Learning, Cycle-Consistent Generative Adversarial Network
Student Number : 2016-30059

Contents

List of Figures

List of Tables

Chapter 1

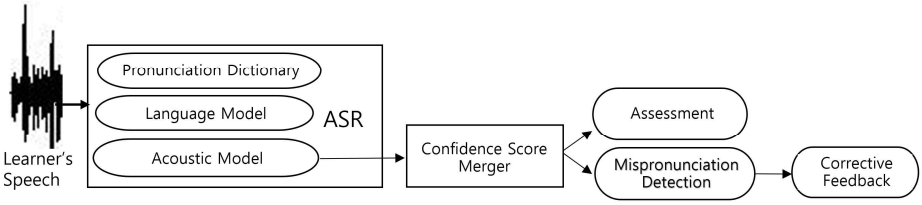
Introduction

1.1. Motivation

Since communication ability is a major purpose of learning a foreign language, precision and fluency in speaking are important goals for language learning. As described in Figure 1, a particular application of Computer-Assisted Language Learning named Computer-Assisted Pronunciation Training (CAPT) uses Automatic Speech Recognition (ASR) technology to assess and detect mispronunciation, pinpointing where pronunciation error occurs, and provide corrective feedback.

CAPT system benefits learners in various aspects and is to be distinguished from conventional pronunciation teaching methods. While a feedback from a native instructor in a classroom environment is conventional to acquire spoken proficiency, the number of available native teachers is too

Figure 1. A conventional CAPT system architecture using ASR technology to automatically assess, detect mispronunciations, and provide corrective feedback.



low to meet each learner's needs. One-to-one lessons by a qualified teacher are ideal to teach pronunciation, compared to the conventional setting in which there are roughly thirty students per teacher (Bloom, 1984). However, only few learners can afford one-to-one tutoring and a complementary approach is to use computer technology to simulate and automate aspects of human one-to-one tutoring. CAPT allows users to follow personalized lessons, at their own pace, and practice as often as they like.

This explains why CAPT systems that assist or substitute human tutors have been attracting considerable attention in recent years (Eskenazi, 2009; Bernstein et al., 2010; Higgins et al., 2011; Zechner et al., 2009; Qian et al., 2010), which motivates the current thesis. As an artificially intelligent tutor, a CAPT software is able to offer individualized tutoring regardless of constraints in time and place, maximizing learning opportunities at learners' convenience.

Moreover, the feedback generation in a CAPT system has an advantage in that it can provide learner's mother tongue (L1)-specifc individualized feedback. The ability to address the L1 specifities is an important advantage because L1 influences the target language production in foreign language

learning, and individualized feedbacks are pedagogically ideal.

Despite the growing popularity in learning Korean as a foreign language and the rapid growth in language learning applications (Chung, 2018), only few researches exist in CAPT systems targeting Korean. Moreover, the existing CAPT applications for Korean tend to be under-researched in terms of non-native speech phenomenon and rely on recording and playback functions instead of exploiting the recent advancement in spoken language processing technologies. The following section surveys the available CAPT systems for English and Korean.

1.1.1. An Overview of CAPT Systems

In recent years, various CAPT systems enabled by dedicated ASR technology have become available. Examples of such systems include "Rosetta Stone" (Rosetta Stone, 2013), "Tell Me More" (Tell Me More, 2013), "EduSpeak" (Franco et al., 2010), "FLUENCY" (Ezkenazi and Hansma, 1998), "ISLE" (Menzel et al., 2000), and "Talk to Me" (Auralog, 2002). "FLUENCY" is an interactive pronunciation trainer with a duration correction module and user adaptive interfaces. Experiments with the application showed that it is dependable and well-accepted by students. "ISLE" uses phone error localization and diagnosis technology, which identifies the mispronunciations and offers detailed error explanations and specifically tailored follow-up exercises that are intended to highlight the contrast between the student's solution and the target, as well as to reinforce the desired articulatory behavior. "Talk to Me" features 3D phonetic animations, pronunciation evaluation, interactive dialogues with progressive learning contents.

For Korean learners of English, CAPT applications have been developed and commercialized such as “Siwon School,” “Speaking Max,” and “Genie Tutor.” “Siwon School” provides conversational speech practices using not only ASR technology, but also multimedia contents, in five different steps: Watch, Speak, Learn, Quiz and Lecture. “Speaking Max” provides Repeat, Lecture, Training, Dictation, Quiz, and Speech exercises based on video interviews of 2,000 native speakers. The Speech exercise includes shadowing of the native speech and answering spontaneous questions within the context of the video. “Genie Tutor” is another dialogue-based application using ASR technology. It provides language proficiency evaluation and accepts free text input, which increases the interactive capability of the system. Although many commercial systems aiming at language learning are available, they tend to be limited in generating feedback that can help learners correct errors (Kim, 2018).

At the research level, CAPT systems have been actively studied with studies focusing on improving the performances in non-native speech recognition, automatic assessment, and mispronunciation detection by using feature engineering and machine learning techniques. Metze (2005) proposed a speech recognition method that adapts to individual speaker and speaking styles using articulatory features in order to improve automatic speech recognition for non-native speech. Automatic assessment technology has been developed in Müller (2010) for oral proficiency that is suited for speakers of South African English, and more generally, in Versant English Test (Berstein et al., 2010) and SpeechRater from ETS (Higgins et al., 2011). These automatic pronunciation scoring have been widely applied in real test conditions in TOEIC® and TOEFL® tests (Zechner et al., 2009). Regarding

mispronunciation detection, Doremalen (2014) developed and tested a CALL (Computer-Assisted Language Learning) application for Dutch as a foreign language based on vowel pronunciation error pattern analysis for non-native speakers of Dutch. Peabody (2006) presented a novel method for representing the acoustic features of vowels to account for non-native variation in vowel production, which are used as the anchoring method for mispronunciation detection system. Since the mispronunciations deviate from the canonical phoneme in many possible directions, recent mispronunciation detection technology uses unsupervised error pattern discovery (Li, 2018).

1.1.2. Survey of Existing Korean CAPT Systems

Existing CAPT systems for Korean are surveyed here with a view to analyze the strength and weaknesses of the status quo, and to define the problem that needs to be addressed. The paragraph below surveys the applications in terms of the knowledge content, and assessment and corrective feedback methods.

"Kmaru SPEECH" provides playback of reference and learners' pronunciation together with the waveform visualization and automatic evaluation. "Learn Korean," "Korean Pronunciation Teaching," and "Pronunciation Practice" all offer playback of reference and learners' pronunciation. "Correct Pronunciation LITE" teaches phonological processes in Korean. However, these applications do not consider Korean phonetic or phonological error pattern analysis in their read speech design. Because of this, the learners cannot tell whether or not their pronunciation was correct, and even if they can hear the difference, it is not able to show the error type.

The automatic assessment methodology in some of these applications use machine learning models built with handcrafted features, such as Goodness-of-Pronunciation scores from the acoustic model and articulation rate (Witt, 1999). However, the remarkable success in deep neural network technology in a variety of machine learning tasks has demonstrated its effectiveness, which could further benefit the performance. We propose two aspects in which the existing methodologies used in CAPT systems can be improved.

First, Deep Neural Network (DNN) approaches in CAPT systems have been only applied and tested in speech recognition, assessment, and mispronunciation detection tasks so far, and to the best of my knowledge, there has been no thorough investigation towards corrective feedback modeling using DNN. Speech correction seems to be a suitable area for testing the capacity of deep learning because beyond the detection tasks, deep learning models are found to provide substantial increases in generation and synthesis abilities. To this end, Chapter 4 in the present thesis introduces a new methodology using Cycle-consistent Generative Adversarial Network (GAN).

Second, the unsupervised end-to-end solution that consists of deep neural network models learns predictive features automatically, which is not desirable in cases when precise corrections are required for learning. Although the end-to-end models may show high performance, a disadvantage occurs when the linguistic knowledge in hand-crafted features cannot be directly utilized. This is problematic in the formation of corrective feedback generation for second language learning, because the type of feature learnt by the neural network model needs to be identified in order to guide the learners what linguistic information needs to be educated. To this end, Chapter 5 in the

present thesis introduces a new methodology using an Auxiliary-Classifier based Cycle-consistent GAN that allows both the unsupervised end-to-end training and an architecture to inject the linguistic knowledge within the network.

1.2. Problem Statement

The goal of the present thesis is to propose a new approach for a Korean CAPT system based on linguistic analysis of L2 Korean speech. It is not the goal of this work to build an entire CAPT system for Korean. Instead, the problem is to first identify pronunciation variation patterns and utilize the knowledge in today's state-of-the-art deep neural network technology for the task of corrective feedback generation.

1.3. Thesis Structure

The thesis can be divided into two parts. The first part comprises of Chapters 2 and 3. Chapter 2 introduces basic concepts of phonology and phonetics and the underlying L2 Korean analysis methods. It discusses these two-level properties of phonemes and phonology and the differences occurring between native Korean and Korean produced by Chinese. Chapter 3 concludes the first part by identifying the linguistic factors affecting the evaluation of L2 Korean, which motivates the use of linguistic features in corrective feedback research later in Chapter 5.

The second part begins with a survey of related works in corrective feedback generation in Chapter 4. By interpreting the problem as a style

transfer, it proposes a method using Cycle-consistent generative adversarial network (CycleGAN) training. This method can serve to generate a corrective speech by unsupervised learning, as the differences predicted in the speaking style, i.e. non-native speech vs. native speech, can be modelled by the generator. In Chapter 5, an auxiliary classifier is trained to model linguistic domain knowledge identified in Chapters 2 and 3. It combines the knowledge with the methodology presented in the second part, proposing a linguistically-motivated auxiliary-classifier CycleGAN.

Chapter 2

Pronunciation Analysis of Korean Produced by Chinese

Pronunciation variations in non-native speech are far more diverse than those observed in native speech. This poses a difficulty for CAPT systems to automatically recognize learners' speech, detect mispronunciations, and provide corrective feedbacks. For an effective CAPT system, it is essential to identify frequent variation patterns based on phonetic and phonological annotations of the non-native Korean speech. Since these human annotations serve as the ground-truth and the ultimate goal is to correct mispronunciations, such annotated corpora is crucial for CAPT system development.

In other L2 languages, many previous studies have analyzed pronunciation variation patterns. Chen (2013) conducted phonetic and tonal error analysis in L2 Chinese produced by 305 speakers of European descent

whose first language is non-tonal. Gut (2009) conducted a corpus-based analysis of phonological and phonetic properties of L2 English and German, and You (2005) studied pronunciation variations of Spanish-accented English spoken by young children using dynamic programming-based transcription alignment on 4,500 words spoken by children. Hong (2014) investigated variation patterns in English segments produced by Korean learners, capturing only the most noticeable segmental variations that were not found in smaller-scale studies.

In L2 Korean, characterization of non-native pronunciation patterns has been limited, and often descriptive and qualitative in nature. These studies described typical non-native Korean speech features observed in classrooms, discussed possible variation patterns using contrastive analysis and language transfer theory, and collected and analyzed speech recordings. Many of them focused on the confusion patterns between the three-way distinction in Korean, which are tense, lenis, and aspirated manners of speech. However, there has not been a thorough study verifying the proposed hypotheses, as such data is scarce, especially for non-native speech in Korean. Moreover, one of the well-known characteristics of Korean speech is its usage of phonological processes, i.e, changes in pronunciation depending on the phonemic context and the part-of-speech of the word. Since many of these phonological processes between syllables may not exist in learners' native languages, variations are likely to occur. Previous studies discussed possible phonological error patterns, but there has not been a follow-up corpus-based analysis.

This Chapter designed a new experiment method and analyzed a large-scale speech corpus of non-native Korean produced by Chinese learners by annotating the pronunciation and phonological variations. Mandarin Chinese

was chosen as the target L1 due to its growing popularity in L2 Korean learning. Statistics as of November 2018 show that 68,184 out of 142,205 foreign students studying in South Korea are Chinese students, which is the largest foreign student group (Chung, 2018). The second largest group of foreign students is 14,614 from Vietnam, followed by 4,358 from Mongolia. The statistics indicate the increasing demand for Korean language learning, especially by learners whose mother tongue is Chinese.

Moreover, Korean produced by Chinese is an interesting topic of research due to the linguistic contrast between the two languages, such as the manner of speech and phonological processes in the target language. These distinctions are not required as the standard pronunciation, and this investigation is expected to provide insights into how we can further develop spoken language technology in Korean targeted for Chinese learners.

2.1. Comparison between Korean and Chinese

The following sections conduct contrastive analyses of phonemes and phonology in Korean and Chinese. Then, it reviews the related works and designs an improved experiment method. Using the improved method, the salient segmental and phonological variation patterns are presented together with discussions on potential research directions that can spawn from the results.

2.1.1. Phonetic and Syllable Structure Comparisons

Table 1. Korean and Chinese Consonants (K: Korean, C: Chinese)

Manner \ Place		Bilabial		Alveolar		Post-alveolar		Palatal		dental		Velar		Glottal	
		K	C	K	C	K	C	K	C	K	C	K	C	K	C
Stop	Unaspirated	p̚	p	t̚	t							k̚			
	Slightly Aspirated	b̥		d̥								g̊			
	Aspirated	pʰ	pʰ	tʰ	tʰ							kʰ			
Affricate	Unaspirated						tʂ	tɕ̚							
	Slightly Aspirated							d̥ʑ							
	Aspirated						tʂʰ	tɕʰ				t			
Fricative	Unaspirated			s	ʂ		s				f		h	h	
	Unaspirated			s̚											
Nasal		m	m	n								ŋ	ŋ		
Liquid				l									l		
Semi-vowel		w						j				w ɥ			
Approximant							ɹ								

Table 2. Korean and Chinese Vowels (K: Korean, C: Chinese)

Height \ Backness	Front		Central		Back	
	K	C	K	C	K	C
Close	I	i y			ɯ u	u
Close-mid	e				o	
Mid				ə		
Open-mid	ɛ					
Open		a	a		ʌ	

Since learners' L1 and L2 both influence L2 production, comparing the phonetic inventories of the two languages helps to predict pronunciation variation patterns. This method, which is called contrastive analysis, has often been criticized for its inadequacy in predicting the mispronunciations in actual learning context. Nevertheless, for L2 phonology, it cannot be denied that contrastive analysis has a predictive power and that it may be able to explain the mispronunciation patterns. The following subsections compare the phonetic inventories, syllable structures, and phonological phenomena in Korean and Mandarin Chinese.

There are 19 Korean consonants excluding the approximants /w,j,ɥ/ (Lee,

1996), and 19 Chinese consonants (Lin, 2007). Chinese language discussed here refers to Mandarin Chinese. The stops and affricates in Korean are grouped into lenis, tense, and aspirated by the manner of articulation, while in Chinese, there are aspirated and unaspirated distinctions. The lenis stops /b̥, d̥, g̊/ and lenis affricate /dʑ̊/ in Korean are slightly aspirated, while the aspirated stops /pʰ, tʰ, kʰ/ and aspirated affricate /tɕʰ/ are heavily aspirated. The tense stops /p͈, t͈, k͈/ and tense affricate /tɕ͈/ are laryngealized and not aspirated. Chinese affricates /tʂ, tʂʰ, ts, tsʰ/ do not exist in the Korean counterpart. Fricatives are grouped as lenis and tense by the manner of articulation in Korean, while they are grouped as aspirated and unaspirated in Chinese. The post-alveolar fricative /ʂ/ and labio-dental fricative /f/ in Chinese do not exist in Korean. In approximants, Korean has the semi-vowels /w,j,ɰ/, which are not individual phonemes in Chinese according to (Lin, 2007). These contrasts are summarized in Table 1.

For vowels, there are eight and five monophthongs in Korean and Chinese, each respectively. The two inventories share /i,u,a/ in common. While /ɯ/ sounds of Korean do not exist in Chinese at phonemic level, /ɛ/ sound of Chinese does not exist in Korean. It should be noted that the scope of contrastive analysis here is at the phonemic level, and an analysis at allophonic level may yield different results. These contrasts are summarized in Table 2.

A syllable in Korean is composed of (C)V(C), a consonant in the onset, a vowel in the nucleus, and a consonant in the coda. The onset and coda consonants are optional. A syllable in Chinese is composed of an optional consonant at the initial and a final, which may be a monophthong or a diphthong, followed by an optional /n/, /ɻ/ or /ŋ/. The differences in syllable

structures show that /n/ and /ŋ/ are the only consonants that can be realized as the syllable coda in Chinese, whereas a Korean syllable allows /g̚, n̚, d̚, l, m̚, b̚, ŋ/ as the syllable coda.

2.1.2. Phonological Comparisons

Phonological processes express systematic phonological sound changes by mapping the underlying representation and the surface level realization, describing how a phoneme stored in the speaker's mind yields what the speaker actually pronounces. For example, in English, intervocalic alveolar flapping occurs when it is placed in between stressed and stressless vowels, which changes the letters ‘t’ or ‘d’ into a quick flap consonant /ɾ/, in words such as ‘butter’ in most dialects of American English. A phonological process that is present in one language may not be present in other languages, which motivates a contrastive analysis of the phonological processes in Korean and Chinese.

In Korean, phonological processes are phonemic changes that occur at syllable boundaries in certain phonemic contexts when producing a sequence of segments, and are described in Table 3. Surface level pronunciations are not always realized as the underlying form, but are directed by these processes. For example, in the word /ʃilla/, whose pronunciation according to its written form is /ʃin̚la/, lateralization rule is applied in the third segment. All rules in the table are required as the standard pronunciation, which means that words that do not conform to these phonological processes may sound foreign or ill-formed.

Table 3. Korean Phonological Processes

Type		Phonological Process	Description
Subsitution	Forward Assimilation	Tensification	When /g̊/, /d̥/, /b̥/, /s/, /dʑ/ are preceded by /p˥,/t˥ or /g˥ they are realized as /k˟/, /t˟/, /p˟/, /s˟/, /tɕ˟/, each respectively
	Back-ward Assimilation	Nasalization	When /p˥,/t˥ or /g˥ are followed by /n/ or /m/, they are assimilated and realized as /n/ or /m/
		Palatalization	The orthographic sequence /t˥ i/ is realized as the sequence /dʒi/ or /tɕʰi/
		Lateralization	The orthographic sequence /n˥ l/ and /ln/ are realized as the sequence /ll/
		Bilabialization	The sequences /n˥ b/, /n˥ p˟/, and /n˥ m/ are realized as /m˥ b/, /m˥ p˟/, and /m˥ m/, each respectively
		Aspiration	When /p˥,/t˥ or /g˥ is followed by /h/, they are realized as /pʰ/, /tʰ/, /kʰ/, each respectively
		Liaison	When a coda segment is followed by a vowel in the next syllable, the coda is resyllabified as the onset of the next syllable
	Other	Neutralization	All coda endings are realized as /k˥,/d˥,/b˥,/n˥,/l/,/m˥, or /ŋ/
Deletion	Consonant cluster simplification		All orthographic double consonants at the coda are simplified as /k˥, /d˥, /b˥, /n˥, /l/, /m˥, or /ŋ/
	/h/ deletion		when placed at the coda, orthographic /h/ is not pronounced
Insertion	Nasal insertion		For the compound nouns with inter slots, /n˥ / is inserted

Table 4. Chinese Phonological Processes

Type		Phonological Rule	Description
Subsit-ution	Backward Assimilation	Bilabialization	The preceding coda is influenced by the following onset segment. For example [Paŋ maŋ] is realized as [pam maŋ] and [nan mien] can be realized as [nam mien].
Deletion	Segment		Deletion occurs when [tʂ lan] is realized as [tʂ la].
	Syllable shortening		In fast speech, two consecutive syllables are integrated into one syllable.
Insertion	Linking segment		for linking segment (Chao, 1968): Nasal or stop sound insertion before a vowel. For example, when [kan a] is pronounced [kan na], a nasal sound is copied to the onset position and pronounced as [kan na].

According to (Lin, 2007), phonological processes in Chinese are not required and may or may not occur. Comparing the Tables 3 and 4 shows that only bilabilization is common in the two languages, while all other phonological phenomena are unique to Korean or Chinese.

For native Korean speakers, the phonological processes are naturally acquired at the beginning of language learning. In contrast, for Korean as L2, in which case grapheme-to-pronunciation education precedes the phonological acquisition, such conversion rules are explicitly learnt. In this vein, it should be noted that the term “phonological process" here may be debatable for L2 phonology because L2 learners do not have the native phonology by its definition, and inevitably use grapheme context information to learn the sound change. Although we considered using the term “grapheme-to-phoneme process," we decided to keep the common English translation “phonological process" in this paper to maintain consistency with the previous researches, such as (Jun, 2018).

2.2. Related works

Kim (2008), Qin (2010), Hwang (2012), Leng (2014) have predicted pronunciation variations using the comparison between Chinese and Korean phoneme inventories, and verified their predictions with learners' utterances. For example, they predicted that variations will occur in lenis stops, since there is no equivalent manner of articulation in Chinese. However, their corpus-based findings differ, as some found confusions occurring between lenis and aspirated stops, while others found confusions occurring between tense and lenis stops. This means that they differ whether the word /k̊i/ (‘air’) has the tendency to be realized as /kʰi/ (‘height’) or /k͈i/ (‘talent’), as 48.3% of lenis stops were realized as aspirated in (Hwang, 2012), while 85.5% of lenis stops were realized as tense stops in (Qin, 2010).

Moreover, they predicted that variations will occur for alveolar stop /d̪ʑ/,

but different results were found whether /ʥadḁ/ ('to sleep') is realized as /tɕʰadḁ/ ('being cold') or /tɕ⁼adḁ/ ('being salty'). For fricatives, different variation patterns were found between /sadḁ/ ('to buy') and /s⁼/ ('cheap'). For liquids, they predicted that flap will have tendency to be realized as /l/ or /ɹ/ due to L1 influence. For example, /naɾa/ ('country') may be pronounced as /nalla/ ('to carry') or /naɹa/. A survey of these studies shows that the related works differ in their variation pattern findings.

The differences in the results can be explained by the amount of data used and the diversity in learner levels. The corpora used in these studies are small in size, as shown in Table 5, which makes it challenging to obtain consistent observations in these analyses. Small-scale datasets are more likely to lead to anecdotal findings that might not easily generalize well to a larger population. Moreover, Kim (2008) and Cho (2013) analyzed the variation patterns of intermediate level learners, while Hwang (2012) and Cui (2002) analyzed those of advanced level learners. It is not surprising to find the differences in analysis results, as the degree of L1 interference would vary across different learner levels.

There are three aspects in which the analysis method in the previous studies can be improved. First, the number of speakers should be larger and balanced in learner levels so that the observations are not dependent on individuals' tendencies. Without such large-scale annotations of speech corpora, it is challenging to characterize the non-native pronunciation patterns to further the understanding in L2 Korean acquisition. To tackle this fundamental challenge, we propose an analysis of a large-scale speech corpus of non-native Korean with detailed human annotations. The corpus size and the level of annotation between the previous and the current study are

compared (Table 5). The corpus used in the present research possesses the following aspects that makes it suitable for studying non-native Korean pronunciation patterns.

- Large in number of speakers.
- Large in number of utterances.
- Large in number of per-speaker data (300 utterances/speaker).
- Complete in coverage of phonemes and phonological processes.
- Diverse in speaker demographic background: gender and learner level balanced.
- Detailed in human annotations, which consists of phonetic and phonological transcriptions.

Second, the learners' variation patterns should be compared with those of native speakers, so that the patterns unique to the learners can be identified. Related works assumed that all deviations from the canonical pronunciation are “errors.” However, deviations from the canonical pronunciation are also found in native speech (Keating, 1998), and the variations that are observed in native speech are unlikely to cause miscommunication (Neri, 2006). Therefore, this study proposes to compare the variation matrices of both Chinese learners and native Korean speakers, in order to capture only the salient variation patterns of the learners.

Furthermore, variation patterns should also be analyzed at the phonological level. Although the previous studies have conducted corpus-based analyses of the segmental variation patterns, no studies have examined the patterns occurring at the phonological level, and this Chapter proposes a two-level analysis both at the segmental and phonological aspects.

Table 5. Comparison of Korean speech corpora produced by Chinese learners. (B: Beginner, I: Intermediate, A: Advanced learner levels, Y: Yes, N: No)

Studies	Speakers		Utterances		Annotation	
	Number	Group	Type	Total	Phonetic	Phonology
Leng (2014)	17	A	38 sentences	646	Y	N
Qin (2010)	46	I	82 words	3,772	Y	N
Kim (2008)	12	I	500 words	6,000	Y	N
Hwang (2012)	20	B,I,A	50 words	1,000	Y	N
Current Study	51	B,I,A	300 words	15,300	Y	Y

2.3. Proposed Analysis Method

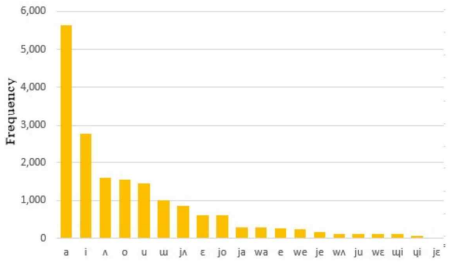
2.3.1. Corpus

This study conducts a larger scale experiment with 51 learners of all levels, in order to find out the learners' prominent variation patterns. L2KSC (L2 Korean Speech Corpus), a speech corpus for Korean as a foreign language is used (Rhee, 2005). The corpus was built to evaluate the acquisition of phonetic and phonological sounds in Korean language by foreign learners of various L1 backgrounds. From L2KSC, this study analyzes read speech of 300 utterances produced by 51 Mandarin Chinese and 51 Korean speakers. For Chinese, there are 17 beginning, 16 intermediate, and 18 advanced level learners, which correspond to the level of class the learner belongs to in a foreign language institute in Yonsei University in Republic of Korea. The read speech script is shown in Appendix I of this thesis.

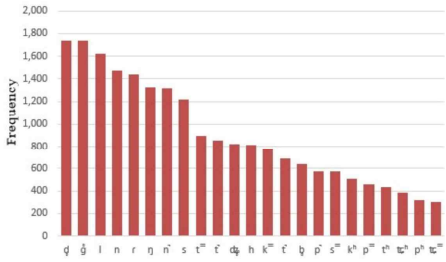
Figure 2 shows the phonetic and phonological distributions in the corpus. In order to be able to separately examine the segmental and phonological

variation patterns, the segments that are affected by the phonological processes listed in Table 3 were manually marked. Figure 3 shows an example of such process. The top row is the pronunciation according to the written form of the characters before the phonological processes are applied, whereas the bottom row is the canonical pronunciation after the correct application of the rules. The horizontal boxes indicate where phonological errors may occur, while all the mispronunciations outside the boxes are counted as segmental errors.

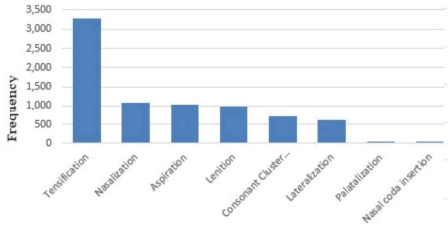
The canonical pronunciation is force-aligned using automatic speech recognition technology. Then, the auditory pronunciation is phonetically transcribed using Korean phonemes and six additional phonemes /tʂ, ʂ, ts, ts^h, f, ɹ/ that are unique to Chinese. All unique Chinese phonemes are added so that L1 interferences can be analyzed, except /y/, whose perceptual difference from /i/ in the same manner and place of articulation in Korean were considered trivial according to the transcribers, and were not perceived as a ‘variation'. When the actual pronunciation is different from the canonical, they were asked to mark the auditory pronunciation. When there is a mismatch in the positions where phonological process occurs, it was considered as a phonological variation, and the error count was increased for the rule. All other mismatches are considered as segmental variations. In order to quantify the relation between the canonical and the actual pronunciation, a confusion matrix is generated.



(a) Consonants



(b) Vowels



(c) Phonological Processes

Figure 2. Segmental and phonological distribution in L2KSC corpus used in this study.

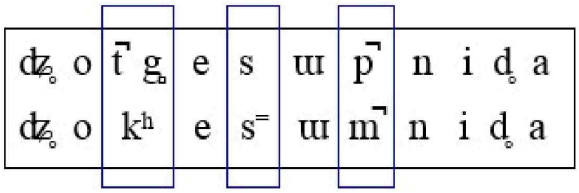


Figure 3. The top row is the pronunciation according to the underlying form of the characters before the phonological processes are applied, whereas the bottom row is the canonical pronunciation after correct application of the process. In this example, which means ‘would like to,’ three segments are affected by phonological processes, shown by the horizontal boxes.

2.3.2. Transcribers and Agreement Rates

The auditory pronunciation is phonetically transcribed by three graduate students with knowledge in Korean phonetics and phonology in the Department of Linguistics in Seoul National University. The transcriber agreement rates are calculated according to the Pairwise Agreement equation.

$$PairwiseAgreement = \frac{No.\ of\ Phonemes\ in\ Agreement}{Total\ No.\ of\ Variations} \times 100 \qquad (1)$$

The agreement rates for Korean produced by Chinese learners and by native speakers are 86.0% and 97.0%, each respectively. Comparing these figures with those of previous studies' rates (Ryu, 2011; Hong, 2014), the reliability of transcription results in this study is verified.

2.4. Salient Pronunciation Variations in Korean Produced by Chinese

2.4.1. Segmental Variation Patterns

The average variation rates for consonants and vowels are 13.74% and 3.35%, respectively. The distribution of the variation rates for different phoneme groups show that the three-way distinctions in lenis, tense, and aspirated stops cover 28% of the errors (Figure 4). This is followed by coda substitution and deletion errors.

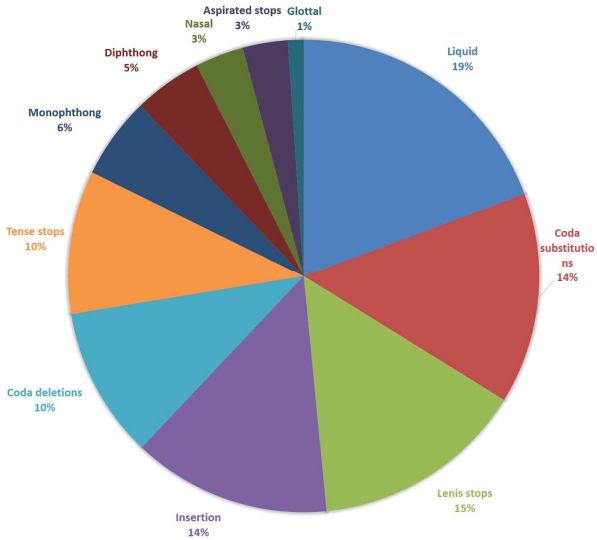


Figure 4. Variation rate distribution for different phoneme groups.

The phonemes with variation rates higher than the average are shown in Table 6. Comparing the “Target Segment” and the “Realized Segment” columns, 33% of flap /r/ are realized as lateral /l/. 20.2% of /p͈/ and 11.9% of /tɕ͈/ are realized as their lenis counterparts, and 7.9% of the lenis affricate /ʤ/ are realized as a Chinese phoneme with a similar place of articulation, /ts/.

Comparing with those of native variations, we can establish the patterns unique to the learners. For example, the flap variations tend to be nasalized for Koreans, while lateralization of flap is unique to Chinese learners' speech. Native variations also occur in coda stops. However, substitutions are common and not deletions, which is unique to the learners. For learners, 20.2% of /p͈/ and 11.9% of /tɕ͈/ are realized as their lenis counterparts, and 7.9% of the lenis affricate are realized as /ts/, a Chinese phoneme with a similar place of articulation. These are not observed in native speech.

Table 6. Salient variation patterns of Korean produced by Chinese. (del.=deletion)

	Target Segment	Count	Variation Rate	Realized Target & Rate (%)					
Consonant	ɾ	1,604	36.10	l	33.0				
	l	4,969	35.10	del.	6.6				
	p˭	942	27.49	b̥	20.2	pʰ	6.9		
	tɕ˭	613	23.49	ts	7.2	tɕʰ	2.9		
	d̥ʑ	1,694	22.67	ts	7.9	tɕ˭	6.9	tɕʰ	3.7
	t˥	1,792	21.94	del.	15.6	k˥	3.8		
	k˭	1,607	21.22	g̊	15.6	kʰ	4.7		
	d̥	3,584	17.75	t˭	14.0	tʰ	3.2		
	t˭	1,840	16.37	d̥	13.1	tʰ	2.8		
	k˥	2,086	14.96	del.	11.0				
Vowel	ɯɿi	230	47.16	e	7.9	ɿi	3.5		
	jɛ	47	25.53	je	19.1				
	wʌ	240	16.31	u	4.7				
	we	474	14.76	ɿi	9.1				
	ja	570	9.12	a	5.1	jʌ	3.7		
	ɿi	143	7.69	we	2.8	u	2.1	ɯɿi	2.1
	ju	239	7.53	ɿi	2.5	u	2.1	jo	2.1
	jʌ	1746	7.1	i	1.8	jo	1.6		
	ʌ	3316	6.12	o	2.8				
	wʌ	573	5.76	we	1.2				
	ɛ	1275	4.94	a	1.5	ʌ	1.1		
	ɯ	2078	4.19	ʌ	1.6				

For vowels, natives also show substitution patterns of diphthongs by monophthongs. Within diphthongs, natives and learners both showed variations with /w/, and variations in /j/ diphthongs were unique to Chinese learners. When these patterns are analyzed at the manner and place of articulation, we find that detensification and coda deletion are the most frequent segmental variations.

Table 7. Comparison between literature survey and corpus-based results on Korean produced Chinese learners

Target Phoneme	Realized Phoneme	
	Literature	Corpus-based (variation rates %)
ɾ	l or ɹ	l (33.0)
p͈	b̥ or pʰ	b̥ (20.2)
tɕ͈	Reference not available	ts (7.2)
l	Reference not available	Deletion (6.6)
d̥ʑ	tɕ͈	tɕʰ (6.9)
t̚	Deletion	Deletion (15.6)
k͈	g̊ or kʰ	g̊ (15.6)
t͈	d̥ or tʰ	d̥ (13.1)
d̥	t͈ or tʰ	t͈ (14.0)
g̊	kʰ or k͈	k͈ (5.38)
b̥	pʰ or p͈	pʰ (5.63)
ɯi	ɯ	e (7.9)
jɛ	Reference not available	je (19.1)
wʌ	Reference not available	u (4.7)
we	Reference not available	ɥi (9.1)
wɛ	Reference not available	ɛ (3.4)
ja	Reference not available	a (5.1)
je	Reference not available	e (4.2)

2.4.1.1. Discussions

Table 7 compares these findings with those in previous works. Regarding the confusions among the three-way distinctions in these studies, tense stops are frequently substituted by lenis stops more than aspirated counterparts. In fact, variations in tense stops were underemphasized in the previous studies, as four of the top ten most frequent variations were tense phonemes. The learners also showed coda deletion patterns, replicating previous studies' results. The findings are consistent with the contrastive analysis hypothesis as flap sounds do not exist in Chinese, and due to L1 influence, they are often

realized as lateral. Other frequent variation patterns include deletion of final consonants, which could be explained by the difference in syllable structures, as /k˺, t˺, p˺,l/ do not exist in Chinese.

Although not listed in the table, new observations were also made in the current results, such as stop insertions and retroflex interferences. These may have been caused by the differences in the syllable structure and mother tongue influence, as retroflex phonemes in L1 do not exist in L2 and syllable structures differ.

2.4.2. Phonological Variation Patterns

Errors were likely to occur in the order of nasal coda insertion, palatalization, lateralization, tensification, nasalization, consonant cluster simplification, and liaison (Table 8). The variation rates tend to decrease as the learner levels increase (Figure 5). The error rates persist for tensification, palatalization, and nasal coda insertions across all learner levels. Some rules are more learnable than others; lateralization and nasalization rules are

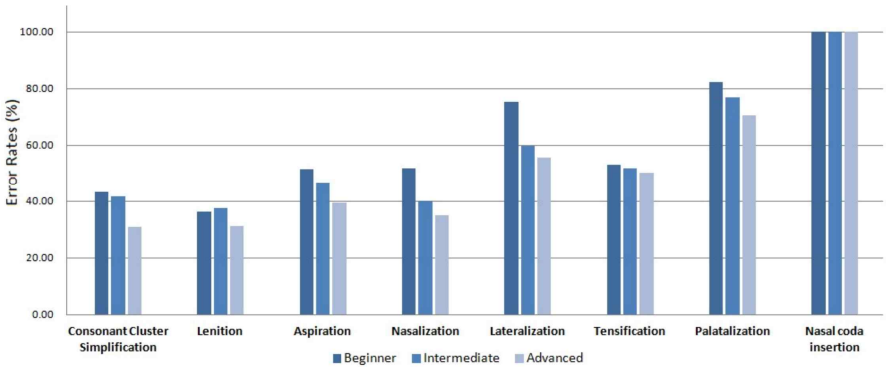


Figure 5. Error rates for each phonological process by learner levels.

Table 8. Frequencies and variation rates of Korean phonological processes in read words produced by Chinese learners. Examples are shown as Canonical Pronunciation > Orthographical Form with English meaning.

Phonological Process	Total Frequency	Error Rate (%)	Examples
Consonant Cluster Simplification	702	38.32	sutak˥ > sutalk˥ (rooster)
Liaison	985	34.72	mogok˥ > mok˥ok˥ (bath)
Aspiration	1032	45.74	dᵊadᵊɯt̪a̪da > dᵊadᵊɯd˥ ha̪da (warm)
Nasalization	1081	42.46	gʌd̪ʑin˥mal > gʌd̪ʑit˥mal (lies)
Lateralization	610	63.93	nallo > nan˥lo (heater)
Tensification	3282	51.68	tɕʰot˥pᵊul > tɕʰot˥p̚ul (candle)
Palatalization	47	76.60	hɛ̞d̪od̪ʒi > hɛ̞d̪od˥i
Nasal coda insertion	47	100.00	namun˥ip˥ > namuip˥ (tree leaf)
Total	7786	47.66	N/A

acquired more easily, as their error rates decrease quickly by 15.4% and 11.6% from the beginner to intermediate groups, each respectively, compared to the 1.60% decrease and 1.23% increase observed for consonant cluster simplification and liaison errors.

2.4.2.1. Discussions

The phonological variations in this study verify the error pattern predictions made by the previous works, which postulated that the Korean phonological processes that are absent in Chinese language are not easily acquired. It also quantifies the level of difficulty for each rule. A question can be then raised; can we explain why lower error rates are observed for consonant cluster simplification, nasalization, and liaison than others? Although these processes are not present in Chinese, it is interesting that

relatively low error rates were observed.

It seems that the positive transfer is a possible explanation for consonant cluster simplification and nasalization. According to the literature survey, backward assimilation exists in Chinese in which the preceding coda is influenced by the following onset. For example, the final /n/ in /nan mien/ is assimilated by the following /m/ and is realized as /nam mien/. When this knowledge in the mother tongue is transferred, Korean nasalization rule can be more easily realized such as in the backward assimilation of pronouncing /apˀmunˀ/ as /ammunˀ/. In this way, the contrastive analysis can explain why some phonological processes are more easily acquired. The same interpretation is possible for the positive transfer of coda deletion in Chinese when realizing consonant cluster simplification in Korean.

However, the contrastive analysis alone cannot explain why low error rates are observed for liaison. Liaison process does not exist in Chinese and contrastive analysis method would predict that since a single syllable forms a single phonological unit in Chinese, negative transfer can occur. However, in contrast, liaison process is acquired relatively easily. A possible explanation is that the high amount of exposure to the process may have resulted in the low error rate. Liaison occurs more frequently than other processes in Korean because there are more phonemic contexts that cause liaison than other processes. It is likely that the learners are more familiar with the more frequent phonological processes than scarcely-observed cases, contributing to lower error rates for liaison than what is expected by the contrastive analysis alone. This may demonstrate a limitation in the predictive ability of contrastive analysis hypothesis, as not every phonological difference necessarily causes a problem when learning a new second language. The

quantified results of the phonological variations not only validate the predictions made by the previous studies, but also provides positive and negative evidence on the effectiveness of using linguistic knowledge and contrastive analysis method for predicting the variations.

2.5. Summary

A corpus-based analysis was conducted with larger number of utterances and balanced speaker levels. By comparing the actual and the reference transcriptions, the variation patterns were analyzed at segmental and phonological levels. Detensification and coda deletion are the most frequent phonetic characteristics, with 36.41% and 58.08% variation rates, while nasal coda insertion and non-realization of palatalization are the most frequent phonological variations. To the best of our knowledge, this is the first large scale corpus-based analysis to have studied the phonological variations in non-native Korean.

Chapter 3

Correlation Analysis of Pronunciation Variations and Human Evaluation

Much research attention has been directed to identify how native speakers perceive non-native speakers' oral accentedness. To investigate the generalizability of previous findings, this chapter examines segmental, phonological, accentual, and temporal correlates of native speakers' evaluation of L2 Korean accentedness. This is a significant topic in language learning, since the results direct how L2 learners can achieve successful oral communication in the language. Indeed, the factors that may interfere with communication, and the degree to which they determine perceptual significance need to be identified for second language instructors, curriculum designers, and language learning software developers, since these standards will be manifested in the corrective feedback systems in CAPT.

For instance, accentedness may not always spill over to lack of intelligibility. In a similar way, whether or not phonological and phonetic errors should be scored with equal weights in Korean, the context in which they cause miscommunication and the degree of listeners' perceptual sensitivity should be investigated. Some types of error influence the overall accentedness score more than others, and it is pedagogically important to obtain a better understanding of the correlation among the factors affecting listeners' perceptions. For an effective CAPT system, it is not only essential to identify frequent variation patterns based on phonetic and phonological annotations of the non-native Korean speech, but it is equally important to identify what actually matters in native speakers' intuitive judgements of accentedness, which measures how much L2 utterances approximate the native speaker norm.

The following section summarizes previous findings on non-native speech assessment and proposes an improved experiment design for L2 Korean, followed by the results.

3.1. Related Works

3.1.1. Criteria used in L2 Speech

In second language (L2) acquisition, a growing number of researchers have emphasized the importance of assessing L2 speech accentedness based on judgments of comprehensibility, accentedness, and intelligibility (Derwing, 1995; Mumro, 1997; Akiyama, 2017; Strik, 2004; Li, 2016; McBride, 2015). They considered L2-dependent factors in English, German, Spanish, Japanese,

Dutch and Chinese. They have studied what kinds of linguistic properties, such as phonetic accuracy, fluency, and grammar errors, are relatively crucial for native speakers' assessment under various task conditions.

For example, according to Derwing and Munro's seminal work on accentedness and native speakers' comprehensibility (Derwing, 1995; Munro, 1997), utterances that are perceived as heavily accented can be highly comprehensible. The finding showed that the degree to which learners approximate the native speaker norm does not necessarily measure how easily L2 utterances are understood. More empirical studies have examined phonological, temporal, lexical or grammatical correlates of L2 German (O'Brien, 2014), Japanese (Akiyama, 2017), Dutch (Strik, 2004), Chinese (Li, 2016) and Spanish (McBride, 2015) comprehensibility.

3.1.2. Criteria used in L2 Korean Speech

The previous studies mentioned above identified the factors affecting speech accentedness for different languages. It still remains open to question whether and to what degree the criteria and the findings in these studies can be generalized to Korean linguistic contexts. The related works considered the L2-specific linguistic characteristics when designing the evaluation criteria, such as tonal realization patterns in Chinese (Li, 2016), and pitch accent in L2 Japanese (Akiyama, 2017). In this section, L2-dependent factors for Korean will be studied in order to identify the linguistic properties of Korean speech that are perceived as crucial for native evaluators.

The evaluation criteria in previous experiments on L2 Korean are surveyed (Table 9). They assessed whether or not meaningful correlations can

be observed between a fixed number of factors and accentedness scores (Jung, 2008; Hong, 2014; Hong, 2016; Lee, 2013; Sayamon, 2016). The accentedness criterion refers to a holistic measure according to the rater's impression of accentedness of an utterance. This holistic measure of accentedness is distinguished from analytic measures since the raters rely on comprehensive impression across the entire utterance rather than paying attention to particular linguistic properties, such as fluency, phonology, or phonetics (Claphaim, 1996).

The filled circles in Table 9 indicate the factors that are highly correlated with the overall accentedness score according to the previous experiment result. There is no filled circle in the columns corresponding to Kim (2017) and Lee (2016) because these studies did not measure correlations between variables, but were interested in longitudinal changes across time. Results in Kim (2017) concluded that fluency score improves for 6 months and starts degrading, while Lee (2016) found that all learners show different improvement patterns over time.

All studies measuring fluency as the evaluation criteria agree that it highly correlates with native listeners' perception of speech accentedness (Jung, 2008; Hong, 2014; Hong, 2016). The correlation was shown to be stronger than segmental accuracy (Jung, 2008; Lee, 2013). However, segmental accuracies, including all the substitutions, deletions, and insertions, are more important according to Hong (2014), and the number of juncture insertion is the most important consideration in Hong (2016).

Table 9: Evaluation criteria used in previous studies assessing non-native Korean speech (○= used as a variable, ●=used as a variable and found to be an important feature).

Evaluation Criteria	(Jung, 2008)	(Hong, 2014)	(Hong, 2016)	(Lee, 2013)	(Sayamon, 2016)	(Kim, 2017)	(Lee, 2016)
Pitch	●						
Juncture	●		●				
Fluency	●			●	●	○	○
Segmental accuracy	○	●		○		○	○
Phonological accuracy	○						
Complexity				○		○	○
Hesitation					○		
Comprehensibility					○		○

The differences in their findings can be explained by the different experimental design. For example, L1-dependent factors may cause disagreements in the correlation tendencies; it may be the case that segmental accuracy is more predictive of the accentedness for L1 Japanese speakers, while suprasegmental factors are more predictive for L1 Chinese speakers. Moreover, for evaluating read speech prompts, accentedness score was used as a criterion (Jung, 2008; Hong, 2014; Hong, 2016), while comprehensibility or complexity measures were used for evaluating spontaneous speech (Lee, 2013; Sayamon, 2016; Kim, 2017; Lee, 2016). Whether the material was read or spontaneous speech would cause differences in the analysis result, as it introduces orthographical influence and knowledge of the vocabulary.

The disagreements above show the need for an improved experiment method that can clarify which linguistic property influences native speakers' judgments of L2 Korean. In this Chapter, we propose to improve the experiment method in two aspects. First, all possible factors related to

accentedness will be included. Some variables in the related works of other L2's have not been considered in the previous L2 Korean studies, which can be a limitation. For instance, no L2 Korean experiments assess the effect of pitch and stress errors, which have been influential factors in other L2 evaluations (Derwing, 1995; Munro, 1997; Akiyama, 2017). The following section introduces an extended coverage of linguistic factor design proposed in this study, and thereby enabling a comprehensive consideration of possible correlations with the accendtedness scores.

Second, we propose to improve the experiment method by including all types of phonological processes, which is one of the characteristics of Korean speech, i.e, changes in pronunciation depending on the phonemic context and the part-of-speech of the word. Several studies have reported that learners of Korean are pronouncing the segments according to their underlying representation, and phonological rules are not realized (Yoo, 2012; Chang, 2014; Chung, 2014; Lee, J, 2005). However, in the correlation studies, only few phonological processes have been included in Jung (2008), and it is necessary to design an experiment that is comprehensive in scope. The extent to which phonological accuracy affects the assessment of L2 Korean speech needs to be thoroughly investigated. The next Section will discuss in more detail what the missing phenomena were in the previous work, and how we propose to improve the experiment.

3.2. Proposed Human Evaluation Method

This section describes the improvements made in the proposed method compared to the previous studies. It elaborates on the reading prompts, variables, evaluation method, speakers, and evaluators.

3.2.1. Reading Prompt Design

Fifty speakers were given 100 sentences to read. The sentences are composed of everyday vocabulary from L2 Korean text books, such as “How many times have you been to Korea?” and “I usually eat dinner when I go home.” The entire script is shown in Appendix II of this thesis. Read speech was used because the canonical pronunciation is predefined, which can be an advantage for discovering error patterns, and also for conducting a research with the beginner level speakers, whose canonical form of the utterance are often impossible to identify.

Moreover, using read speech prompt enables a comprehensive analysis of phonological accuracy. For this study, 50 sentences were composed of phonological processes that are balanced in number and types. In total, there are 264 instances of phonological processes in the prompt (Table 10), including five common phonological processes occurring both cross and within-morphemes. For example, tensification rule in the word ‘worry’ (gʌk̚z̿ʌŋ) occurs within morpheme, whereas the aspiration rule in the word ‘would like’ (ʤ̬okʰes̿ɯm̚niɾ̥a) occurs across morpheme. Regarding sentence types, the 50 sentences consist of 42 statements and 8 questions.

Table 10: Frequencies of phonological processes occurring in the fifty sentences used in this experiment.

Phonological Process	Cross-Morpheme	Within-Morpheme
Liaison	82	5
Tensification	33	13
Nasalization	10	25
Aspiration	26	4
Palatalization	2	3
Total	158	106

3.2.2. Evaluation Criteria Design

The purpose of the current investigation is to examine the generalizability of previous findings (Jung, 2008; Hong, 2014; Hong, 2016; Lee, 2013; Sayamon, 2016) and resolve the disagreements in their results. Since phonological processes were included as the L2-specific characteristic, we also examine whether and to what degree the correct realization of phonological processes affects L2 Korean perception. The following five variables have been defined as the evaluation criteria: segmental accuracy, phonological accuracy, prosody, fluency, and holistic impression of accentedness. Upon listening to each sample, the raters used 1-5 Likert scale (5: perfect, 4: good, 3: acceptable, 2: poor, and 1: very poor) to evaluate.

- Accentedness: As employed in previous studies (Jung, 2008; Hong, 2014; Hong, 2016), accentedness was rated by the evaluators' impressionistic and holistic judgments of the overall utterance in the scale of 1 to 5, without paying attention to

specific linguistic features. For example, even if a part of an utterance digresses from the canonical, they can assign high scores if it is perceived as acceptable.

- Fluency: The evaluators rated fluency based on rate of speech, juncture, pause, and filled pause. For example, novice learners tend to speak slowly and pronounce each syllable separately, which is not observed in native speech and such instances would discount the fluency score.
- Phonological accuracy: All syllables where phonological rules occur and their types are marked in advance for the five different phonological phenomenon. In this way, the raters know which errors to listen to. They counted the number of errors and gave scores based on the error rate.
- Segmental accuracy: The raters phonetically transcribed all segments and rated according to the rate of mismatch between the canonical and realized pronunciations.
- Prosodic accuracy: The raters judged the appropriateness of prosody realized at lexical and sentence levels. For example, if a question is perceived as a statement due to an inappropriate pitch realization, the utterance will receive a low score.

With the criteria design, the present investigation method improves the previous methods in the following three aspects. First, it covers a wider range of evaluation criteria, compared to the two or three variables (Hong, 2014; Hong, 2016; Lee, 2013; Sayamon, 2016), to five variables. Since the raters had a prior knowledge of the read prompts and therefore, comprehensibility

and grammar accuracy were not included as a variable. That is, the degree of effort required by the raters to understand an utterance could not be independently measured in this study by the nature of read speech task. This is also consistent with the previous studies mentioned in Section 3.1.1. that evaluated read speech by holistic impression of accentedness instead of comprehensibility.

Second, within the phonological accuracy, the sentences comprise of higher diversity. Aspiration and palatalization error types have been added, in addition to lenition, nasalization, and tensification. This is important because phonological processes are pedagogically meaningful where the learners may need explicit instruction.

Moreover, the experiment is conducted with speakers from more diverse L1 backgrounds compared to the related works in L2 Korean. It is possible that the correlation studies showed differing results because of the diversity in speakers' backgrounds. In order to reduce the disagreements arising from L1 effect and gain a better view of the overall tendencies, we designed the experiment to include speakers of more diverse backgrounds, including Mandarin Chinese, Japanese, Cambodian, Vietnamese, and Filipino.

Note that only one pronunciation per word was defined as the canonical form when evaluating the segmental accuracy. There are certainly regional variations that are also recognized as acceptable pronunciations in standard Korean, which means that some may not be perceived as an error. However, predefined standard Korean pronunciation exists according to the National Institute of Korean Language, and is the form of Korean that is accepted as a norm. Considering that the purpose of current research is a pedagogical application, it was desirable to keep the correct reference as the gold standard.

Therefore, multiple correct answers were not allowed in this experiment. In addition, there is no consensus on what counts as an 'acceptable variation,' and would cause a confusion in the scoring process.

3.2.3. Raters and Agreement Rates

Each utterance was scored by four native Korean graduate students in Seoul National University with knowledge in Korean phonetics and phonology. Since phonological accuracy was included in the evaluation criteria, it was necessary to recruit raters with detailed knowledge in Korean phonological rules in this evaluation task.

The evaluators practiced scoring with the established guidelines to ensure inter-rater consistency. Before the four raters could officially start scoring, we made sure that the inter-rater correlation in Cronbach's alpha was at least 0.6 on the first 50 utterances for training purposes. With a view to utilizing the material for developing an automatic scoring model in a CAPT in a future study, it was desirable to obtain consistency in scoring. Biweekly training and discussion sessions were held for monitoring inter-rater consistency throughout the scoring and annotation period, which took about five months.

The four raters demonstrated general agreement (α = 0.88) on the accentedness rating task over 2,500 utterances, suggesting that they share similar intuitive notion of what it meant by holistic impression of accentedness in L2 Korean speech. The coefficients reported in the previous studies confirm that the results are reliable (α = 0.82 (Saito, 2017), 0.88 (Hong, 2014), 0.89 (Hong, 2016), 0.74 (Sayamon, 2016)).

Table11. Accentedness score distribution for 2,500 utterances, each rated by four native speakers.

Score	1	2	3	4	5	Total
No. of Utterances	1,273	2,906	3,388	1,971	462	10,000

3.3. Linguistic Factors Affecting L2 Korean Accentedness

A set of correlation analyses was conducted to examine how accentedness ratings were related to the four linguistic variables defined in the previous section. The mean and standard deviation of accentedness scores are 2.94 and 0.98, respectively, and their distribution is summarized in Table 11. In the following analyses, all raters' scores were averaged to derive a single score for the perceived accentedness of each utterance.

3.3.1. Pearson's Correlation Analysis

All variables are strongly correlated with accentedness scores (Figure 6). Among the variables, accentedness was most strongly correlated with

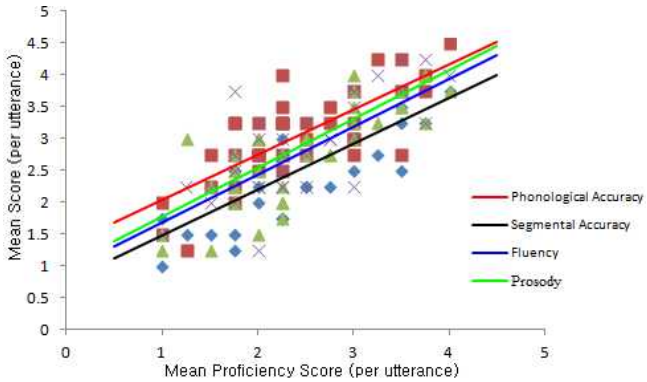


Figure 6. Linguistic correlation with Accentedness scores according to Pearson measure. Correlation is the highest in the order of segmental accuracy (r= 0.81), fluency (r= 0.80), prosody (r= 0.76), and phonological accuracy (r= 0.74).

segmental accuracy (r=0.81) and fluency (r=0.80), and relatively weakly correlated with prosody (r=0.76) and phonological accuracy (r=0.74). All correlations are statistically significant (p<0.0001).

3.3.2. Discussions

The findings indicate that speech with higher accentedness ratings was comprised of fewer segmental errors, and was fluently spoken with an appropriate rate of speech. This suggests that the raters similarly relied on segmental and fluency information during their accentedness judgments.

Similar to the correlation analyses in previous studies, we confirm that speech rate is a significant predictor of speech accentedness. However, it is a new finding in this study that segmental accuracy is an even better predictor

of accentedness than speech rate. Although the difference between them is small, the finding is significant because most previous studies with L2 Korean did not consider segmental accuracy to be the most important measure of accentedness.

Another important finding in this experiment is that prosody is also correlated with accentedness (r = 0.76). Most studies with L2 Korean did not include prosody as a variable, which can be partially explained by the fact that Korean is a syllable-timed language and it is easy to assume that the difference in prosody would not be perceptually significant. However, the experimental result in the current study is contrary to the expectation. In fact, prosody is shown to have an even higher correlation than phonological accuracy, whose average error rate was higher than those of segmental accuracy. according to Chapter 2.

One possible explanation is that mispronunciations at the phonological level do not always spill over to semantic confusion as much as segmental mispronunciations do. For example, pronouncing the word /tɕʰukʰa/, which means ‘congratulations’, as the underlying text form /tɕʰukʰa/ without applying the aspiration rule does not change the word into a different meaning. In contrast, a segmental error pronouncing /nalla/ as /naɾa/ would cause a semantic change from ‘to carry’ to ‘country,’ and pronouncing /p̚ang/ as /pʰang/ would cause the change from meaning ‘bread’ to ‘room.’ Therefore, it may have been easier for the raters to be stricter when judging the accentedness of an utterance with a segmental mispronunciation than those with phonolgical mispronunciation, resulting in the higher correlation scores.

3.3.3. Implications for Automatic Feedback Generation

The results indicate that the effects of segmental accuracy, fluency, prosody, and phonological accuracy are all positively correlated with L2 speakers' oral proficiency scores obtained from native listeners' judgements, in the order of their importances. Using this conclusion in relation to Chapter 2 findings directs which phenomenon deserve higher priority in the corrective feedback. For consonants, six out of the ten most salient variation patterns were confusions between tense and lenis stops. For vowels, nine out of twelve most salient variation patterns were confusions among diphthongs and monophthongs. Therefore, the results of this study can be used to suggest not only that teaching accurate articulation is important in L2 Korean speech, but also that within the phonetic inventory, the manner of articulation of tense and lenis for consonants, and of mophthongs and diphthongs for vowels are important.

Another important finding in this experiment is that prosody is also correlated with accentedness. Most studies with L2 Korean did not include prosody as a variable, which were partially explained by the fact that Korean is a syllable-timed language and it is easy to assume that a prosody error would not spill over to semantically discriminative properties. Both segmental and prosody accuracies, however, need to be considered with an importance in the feedback generation system for L2 Korean.

3.4. Summary

Since certain error types have more perceptual importance than others, it is necessary to discuss the types of error that deserves more importance than others, which motivates the experiment in this Chapter. Using the speech produced by fifty L2 learners of Korean from five L1 backgrounds, the linguistic correlates of accentedness were identified. According to the results in the correlational analysis, our findings were generally consistent with the pervious literature, in that fluency score is a good measure of oral proficiency, including speech rate, juncture, and other temporal features.

The new finding in this study is that segmental accuracy demonstrates the highest correlation with accentedness. Moreover, native listeners are more sensitive to prosody than it was predicted, and may indicate this factor deserves more attention in L2 Korean learning. The results in this Chapter will be utilized in the automatic feedback generation system later in Chapter 5.

Chapter 4

Corrective Feedback Generation for CAPT

4.1. Related Works

Feedback is a critical component in pronunciation training. For the learners to speak L2 fluently and accurately, it is important they practice speaking and receive an appropriate feedback. In traditional classroom settings, there is generally not enough time for sufficient practice and feedback on speaking performance, and a CAPT system is often used to automatically diagnose mispronunciations and offer a corrective feedback for pronunciation training; generating a corrective feedback is an important issue in the area of spoken language technology for education.

In the existing Korean CAPT systems with a corrective feedback

function, according to the survey in Chapter 1, the native pronunciation of the word is recorded in advance and played to the learners. The process of recording the reference pronunciation by a human teacher for each word may be painstaking. With a view to improve the existing methodology, we survey automatic methods in corrective feedback generation in the following sections.

4.1.1. Prosody Transplantation

The traditional feedback way of CAPT is to show the pairwise differences between the target speech and the mispronunciation in various aspects, like the speech wave, speech formant, and articulation (Kawahara, 2002). A number of research efforts have been made to transform foreign-accented speech into its native-accented counterpart. In previous works, speech conversion methods for pronunciation teaching have been studied for Korean and Japanese learners of English, Italian learners of German, Japanese learners of Italian, and for English learners of Mandarin Chinese (Yoon, 2007; Ozawa, 1990; Tillmann, 2006; Debora, 2015; Seneff, 2006). These studies were based on the prosodic transplantation technique (Vitale, 2012), using PSOLA (Pitch-Synchronous Overlap and Add) algorithm (Moulines, 1989). Through this technique, the acoustic parameters including pitch, intensity, articulation rate, and duration of the native speakers are transferred to the learners' speech. It allows the manipulation of prosodic cues while keeping the segmental dimension intact, such as prosody transplantation.

Time-domain PSOLA is most commonly used due to its computational efficiency (Kortekaas et al. 1997). The algorithm consists of three steps: the analysis step where the original speech signal is first divided into separate but

often overlapping short-term analysis signals, the modification of each analysis signal to synthesis signal, and the synthesis step where these segments are recombined by means of overlap-adding (Charpentier et al. 1989; Valbret et. al 1991). Short term signals are obtained from digital speech waveform by multiplying the signal by a sequence of pitch-synchronous analysis window $h_m(n)$:

$$x_m(n) = h_m(t_m - n)x(n). \quad (1)$$

where t_m is the pitch-mark and *x(n)* consists of a sequence of short-term signals $x_m(n)$. The windows, $h_m(n)$, which are usually Hanning type, are centered around the successive instants pitch-marks. These marks are set at a pitch-synchronous rate on the voiced parts of the signal and at a constant rate on the unvoiced parts. The used window length is proportional to local pitch period and the window factor is usually from 2 to 4 (Charpentier 1989; Kleijn et al. 1998). The pitch markers are determined either by manually inspecting the speech signal or automatically by pitch estimation methods (Kortekaas, 1997). The segment recombination in synthesis step is performed after defining a new pitch-mark sequence.

As the manipulation of fundamental frequency is achieved by changing the time intervals between pitch markers and those of duration is achieved by either repeating or omitting speech segments, the application is that prosodic aspects of a native speaker can be imposed on non-native segments, and vice versa. This makes it possible to maintain intelligible signals while selectively manipulating prosodic cues.

The algorithm used for foreign accent transplantation has also been referred to as ‘prosody cloning’ (Yoon, 2007) or ‘prosodic transplantation’ (Gili Fivela, 2012). First of all, the method requires at least two sentences, one

produced by a native speaker and one by a non-native speaker. The transplantation of prosody can be applied using a signal manipulation software, such as Praat (Boersma and Weenink, 2013). It is then possible to automatically superimpose the duration and f0 of one sentence on the segments of the other. The segments of the recipient sentence are first stretched or shrunk in order to match the duration of the donor sentence, and then the f0 contour of the donor sentence is superimposed on the recipient segments. This method has been established and adopted as a method for foreign accent rating in several experimental studies published throughout the last decade, to rank the importance of the prosodic cues involved in foreign accent perception.

Since speech intelligibility is affected by both prosodic and segmental errors, it is beneficial to also achieve segmental transplantations. In Felps et al. (2009), Frequency Domain-PSOLA was employed to replace the spectral envelope of the learner with that of the normalized native speech to achieve the segmental transformation. In this method, the learner's spectra were flattened and multiplied by the native speakers' envelope. In order to reduce speaker-dependent information in the teacher's spectral envelope, Vocal Tract Length Normalization was performed using a piecewise linear function.

4.1.2. Recent Speech Conversion Methods

PSOLA method heavily relies on feature extraction, such as pitch, duration, and spectral envelop extractions of both native and non-native speech, and vocal tract length normalization for each speaker, to mention a few, which require complex pipelines consisting of domain-specific or fine-

tuned techniques.

More recent work has also addressed speech conversion using deep neural network and an end-to-end architecture that directly generates the speech representation without the feature engineering process (Bearman, 2017). Haque (2018) uses an end-to-end model, conditioned on speaker identities, to transform word segments from multiple speakers into multiple target voices. Biadsy (2019) introduced an end-to-end-trained speech-to-speech conversion model that maps an input spectrogram directly to another spectrogram. The network is composed of an encoder, spectrogram and phoneme decoders, followed by a vocoder to synthesize a time-domain waveform. These models succeeded in word and pitch-level transformations of the voice, many-to-one voice normalization, and atypical speech normalization. However, the speaker identity is lost in these approaches. In the section 4.2., we propose a novel feedback generation method.

4.1.3. Evaluation of Corrective Feedback

A handful of studies have suggested methods to evaluate the goodness and the effectiveness of self-imitative feedback. Those that measures the goodness consider the linguistic and technological aspects, paying attention to individual phonemes and sound quality, while those that evaluate effectiveness measure pedagogical value in corrective feedback. For example, four pedagogically critical criteria for feedback in CAPT was prescribed (Hansen, 2006); a feedback should be easy to understand (comprehensible), a feedback should determine if the correct phoneme was used with the correct length, and a feedback should suggest actions for improvement (corrective).

Other studies on automatic speech conversions designed a perceptual protocol to evaluate the effectiveness of the method along three dimensions: foreign accentedness, speaker identity, and signal quality (Felps et al., 2009). More recent studies evaluated whether the converted speech preserves the linguistic content of the original input signal by reporting the word error rate as a measure of intelligibility (Biadsy, 2019). They also reported the mean opinion score (MOS) on the naturalness, voice similarity, accentedness, background noise and disfluencies. The survey on evaluation criteria employed in previous studies shows that perceptual tests on accentedness and sound quality may be used to validate the speech conversion performance in a CAPT system.

4.2. Proposed Method: Corrective Feedback as a Style Transfer

We begin by asking the question: what constitutes a foreign accented speech? A foreign accent can be defined as deviations from the expected acoustic and prosodic norms of a language. The type of deviations is influenced by the context of the speech, including the speaker's mother tongue background, the sentences or words before and after the utterance, speaker's intention, whom the speaker is addressing, and the speaker's environment, to mention a few. A foreign accented speech, or a certain style of speech, is necessarily influenced by its context. All kinds of deviations can be understood as a particular style, and given a linguistic content for example, I could imagine a change in the speaking style in my voice from addressing my

advisor in a meeting, to teaching a group of students in a classroom. Also, what if I, with a Korean mother tongue, were to impersonate a British-accented English? A brief listening to such accent makes it possible to imagine how I would have rendered such sentence: perhaps pronouncing certain vowels and the letter “r” differently. Keeping this accented speech, I could also speak the sentence while imitating an actress’ voice that I know.

It is possible to imagine the sound despite never having seen a side by side example of my impersonated speech next to a British-accented English. Instead, I can learn the style of British-accented English speeches by listening to speech samples and use my knowledge of the characteristics. We can learn about the stylistic differences between the speeches, and thereby imagine what the speech might sound like if we were to “translate” it from one set into the other.

Recently, GANs (Generative Adversarial Networks) have shown promising results in image style transfer and researchers have investigated this problem extensively. The problem can be posed as translating an input image into a corresponding output image; a scene can be translated into another style, rendered as an RGB image, a gradient field, an edge map, a semantic label map, etc. The image community has already taken significant steps in this direction.

We argue that the style transfer method is also capable of learning a style of speech and of transferring the style to another domain. Motivated by the recent successes in GAN’s ability in the style transfer problem, the current thesis adopts GAN to convert foreign-accented speech from a non-native speaker into fluent speech with a native accent. The subsections below further explain the potential advantages of the proposed method.

4.2.1. Speech Analysis at Spectral Domain

One way to analyze a speech is by examining their spectrograms, which visually represent the varying short-term amplitude spectra of the speech waveform. Spectrograms remain a dominant acoustic representation for both phoneme and word-level tasks. It carries phonetic information, and inspired by the process where a human expert "read" a spectrogram, the practice of using this knowledge for speech recognition tasks is common in the discriminative setting (Hershey et al., 2017). The machine can be taught on which cues to focus on in order to identify and learn segmental and supra-segmental information in the graphical representation of speech.

Being able to identify a speech in the spectral domain also suggests that native and non-native speech differences are present in the spectral analysis. This is confirmed by comparing the differences between native and non-native spectrogram pairs of the same utterances. Figure 7 shows an example of a spectrogram pair for the word "half a year." While the left spectrogram captures the resonances of the vocal tract during a diphthong articulation, the right spectrogram shows its monophthong version. As a consequence, the two spectrograms can be differentiated by the number and movements of the darkness bands, showing that non-native speeches are more likely to substitute diphthongs by monophthongs than the native speech. By observing more spectrogram examples, we obtain linguistic differences including final stop deletions, exhibited by the voiced and unvoiced region contrasts in the spectrograms, and lenition of tense consonants, which is demonstrated by the voice onset time in the spectrograms. Moreover, the presence of rhotic vowels

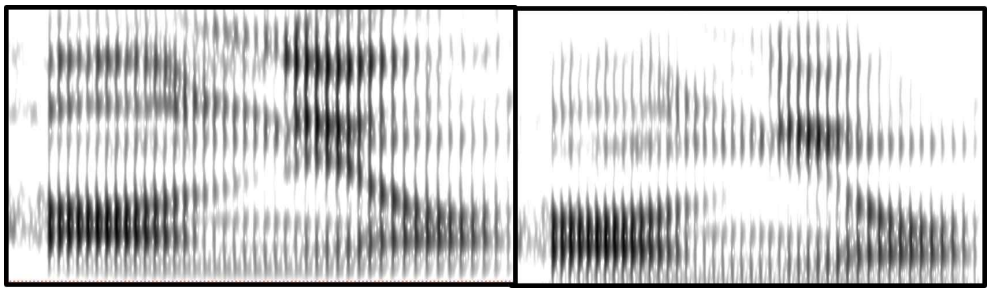


Figure 7. An example of a spectrogram pair for the word “half a year (반년)” in Korean uttered by a native (left) and foreign-accented (right) speakers. The spectrogram comparison is able to capture linguistic similarity and differences, which motivates the idea of using CycleGAN.

in the formant frequencies of the non-native spectrograms is not observed for native counterpart, as the sound does not exist in its phonetic inventory. At the suprasegmental level, the articulation rate and total duration of the native speakers tend to be shorter than the learners’ speech. These findings can be confirmed by analyses of the auditory variation patterns in Yang and Chung (2015).

These observations indicate that spectrograms contain rich information that is enough to differentiate the characteristics of native and foreign-accented utterances in linguistic domains. This motivates the idea for a spectrogram learning using image-generating GAN, where the latent space in the audio of non-native linguistic domain is mapped to that of native linguistic domain.

The examination at the spectral domain also motivates the idea to introduce cycle consistency loss in GAN. Despite the differences between the two domains, non-native and native, we also find that they share an

underlying structure. Since the goal of this work is to generate only speech sounds and not an arbitrary audio, we hypothesize that training the network to learn a high level representation of the underlying language serves to bias the spectrogram decoder predictions toward a representation of the same underlying speech content. Assuming that there is some underlying spectral structure shared between non-native and native linguistic domains, the trained network can be thought of as learning a latent representation of the input that maintains information about the underlying linguistic content.

The current thesis attempts to accomplish this by adopting CycleGAN, which enforces forward-backward consistency between two different domains and is expected to be an effective way to regularize such structured data (Kalal et al., 2010). This motivates the proposal in the current thesis to experiment with CycleGAN. The formulation of the algorithms will be shown in a later section.

4.2.2. Self-imitative Learning

A possible advantage of interpreting the feedback generation problem as a style transfer is that it allows self-imitative learning. In self-imitative learning, the characteristics in native utterances are extracted and transplanted onto the learner's speech. Listening to the manipulated speech enables students to compare the differences between the accented utterances and the native counterparts, both in their own voices, and to produce native-like utterances by self-imitation. The rational of self-imitating feedback is that, by stripping away information that is only related to the teacher's voice quality,

the students can perceive differences between their accented utterance and their ideal accent-free counterparts.

The pedagogical benefit of self-imitating learning is that it provides a form of feedback that is implicit, corrective, and encouraging. A handful of studies have suggested that it would be beneficial for L2 students to be able to listen to their own voices producing native-accented utterances (Felps, 2009). Studies in CAPT also found that the better the match between the learners' and native speakers' voices, the more positive the impact on pronunciation training (Gutierrez-Osuna, 2009), emphasizing the importance of the student and teacher voice similarity for the enhancement of pronunciation skills.

As studies have hypothesized that self-imitative feedback is encouraging and effective, they also evaluated the method with an experiment. For example, one group of students was trained to mimic utterances from a reference English speaker, whereas a second group was trained to mimic utterances of their own voices, previously modified to match the prosody of the reference English speaker (Nagano and Ozawa, 1990). Pre- and post-training utterances from both groups of students were evaluated by native English listeners. Post-training utterances from the second group of students were rated as more native-like than those from the first group.

More recently, the relationship between the student/teacher voice similarity and pronunciation improvement was investigated (Probst, 2002). Several teacher voices of the same sentence were recorded in advance and were played to the students as a corrective feedback. Results showed that learners who imitated a well-matched speaker improved their pronunciation more than those who imitated a poor match. Consistent with the findings, a few CAPT tools have begun to incorporate prosodic-conversion capabilities.

These tools allow L2 learners to re-synthesize their own utterances with a native prosody through a manual editing procedure (Martin, 2004).

The studies discussed above indicate that the learner's own voice with corrective prosody is more effective than prerecorded utterances from a native speaker. Assuming that a foreign-accented style of speech can be learnt in the GAN architecture, we hypothesize that the method will be capable of generating a corrective feedback that is self-imitative.

4.2.3. An Analogy: CAPT System and GAN Architecture

The previous sections explained the motivations and possible advantages of using CycleGAN. There is also a higher-level and yet, practical motivation for the proposal with respect to the ultimate goal of building a CAPT system.

Figure 8 compares the two CAPT system architectures; the traditional and the GAN-based system. The traditional architecture is the same as Figure 1 in Introduction of the present thesis. The proposed GAN-based CAPT system has three advantages. First, the proposed architecture, thanks to the adversarial nature of GANs, connects speech assessment and corrective feedback into a single network. While the generator outputs a native speech feedback. the discriminator's confidence score on the native-likeness of the generated spectrogram can be translated into an assessment score in a CAPT application. One of the difficulties in implementing a CAPT system is the integration of independent modules into a single architecture. The proposed adversarial structure of GAN incorporates these individual tasks in a single network, and thereby improving the connectivity and efficiency.

Second, the traditional architecture relies on ASR performance. For

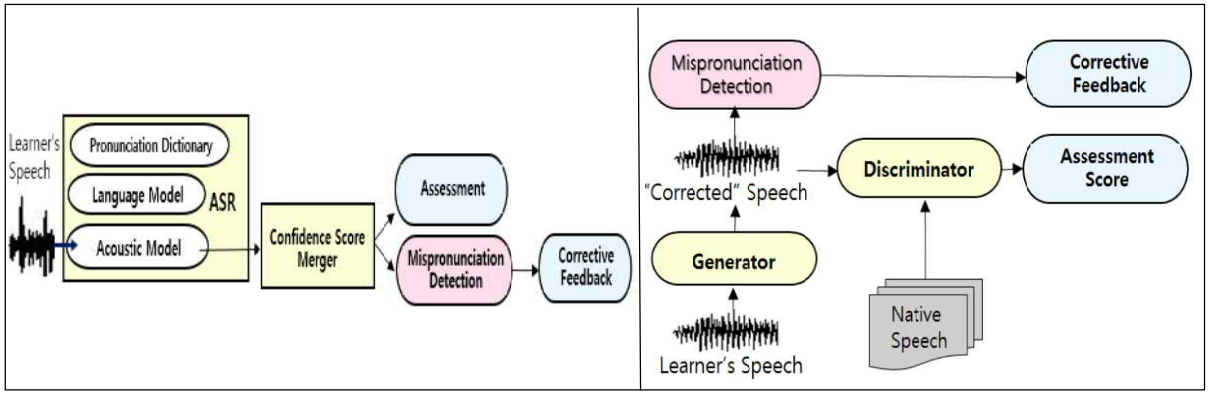


Figure 8. Traditional (left) and proposed (right) CAPT system architecture using GAN algorithm to automatically assess and detect learner's mispronunciation and provide corrective feedback. The discriminative and generative abilities in GAN may be both exploited as corrective feedback and assessment components in the potential CAPT system.

example, errors in the acoustic model can result in an incorrect confidence score, and hence propagating the error to assessment and detection models. In contrast, the proposed method does not rely on the acoustic model performance, making the assessment and feedback performance more directly controllable.

The third advantage is that the feedback and assessment models in the proposed system are end-to-end without the traditional feature extraction processes. The traditional automatic speech assessment software uses combination of 29 or more features (Higgins et al., 2011) in order to predict the score. In contrast, features are learnt automatically in an end-to-end method, which allows easier language expansion of the CAPT system with various L1 and L2 combinations.

4.3. Generative Adversarial Networks

GANs have attracted attention for their ability to generate convincing images and speeches. The advantage of using GANs for style transfer is that the model learns a loss function for scoring the quality of the results automatically, compared to manually designing effective losses.

Gatys (2016) studied artistic style transfer, combining the content of one image with the style of another. In order to transfer photographic style, Luan (2017) added semantic segmentation as an optional guidance and imposed a photorealism constraint in the transformation. Taigman (2016) adopted GAN and variational autoencoder as the mapping function to enforce the transformed output to be similar to the source. Isola (2017) explored GANs in the conditional setting, in which the generator is conditioned on a given image

in the target domain. Zhu (2017) introduced CycleGAN which uses generative network together with a cycle consistency loss to encourage the distribution of the mapped images to be indistinguishable from that of the real images in the target domain. Chang (2019) introduced variants of cycle consistency losses as asymmetric functions to ensure the successful transfer high frequency details.

GANs (Goodfellow, 2014) are generative models that learn a loss that tries to classify if the output image is real or fake, while simultaneously training a generative model to minimize this loss. This adversarial learning process is formulated as a minimax game between G and D, which is formulated as:

$$\min_{G} \max_{D} \quad V(D,G) = \mathbb{E}_{x \sim P\mathrm{data}(x)}(\log D(x)) + \mathbb{E}_{z \sim Pz(z)}(\log (1-(G(z))) \qquad (2).$$

where *Pdata(x)* is the real data distribution, and *Pz(z)* is the prior distribution. For a given x, *D(x)* is the probability x is drawn from *Pdata(x),* and *D(G(z))* is the probability that the generated distribution is drawn from *Pz(z)*.

The generator (G) maps the training samples to samples with a prior distribution by imitating the real data distribution to generate fake samples. G learns the mapping by means of an adversarial training, where the discriminator (D) classifies whether the input is a fake G sample generated by G or a real sample. The task for D is to correctly identify the real samples as real, and thereby distinguishing them from the fake samples. The adversarial characteristic is due to the fact that G has to imitate better samples in order to make D misclassify them as real samples. The misclassification loss is used for further improvement of the generator. During the training process, D back-

propagates fake samples from G and correctly classifies them as fake, and in turn, G tries to generate better imitations by adapting its parameters towards the real data distribution in the training data. In this way, D transmits information to G on what is real and what is fake.

In the following two subsections, two variants of GAN used in the experiment will be introduced.

4.3.1. Conditional GAN

Conditional GANs (cGANs) translate an image from the source domain to the target domain conditioned on a given image in the target domain. It requires that the generated image should inherit some domain-specific features of the conditional image from the target domain (Isola, 2017). This makes cGANs suitable for image-to-image translation tasks, where we condition on an input image and generate a corresponding output image. In this setting, G tries to minimize the objective against an adversarial D that tries to maximize it, with the following objective function:

$$L_{cGAN}(\mathrm{G},\mathrm{D}) = \mathbb{E}_{x,y}(\log D(x,y))\mathbb{E}_{x,y}(\log(1\text{-}D(x,G(x,z))). \quad (3)$$

Isola (2017) demonstrated that cGANs can solve a wide variety of problems by testing the method on nine different graphics and vision tasks, such as a map to satellite image transfer and a product photo generation from a sketch.

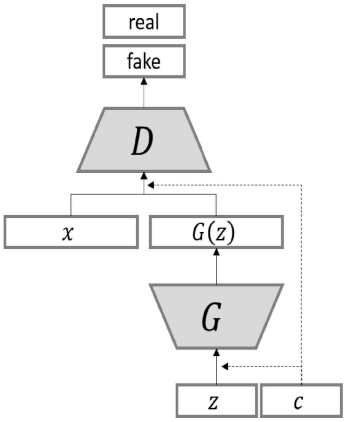


Figure 9. Conditional GAN architecture. The discriminator and the generator are conditioned on c, an additional input layer with values. The added vector of features guide G to figure out what to do.

4.3.2. CycleGAN

The adversarial loss alone may not guarantee that the learned function can map the input to the desired output. In our case, this may result in wrong corrective feedbacks, which would be highly undesirable for feedback generation in CAPT. Zhu (2017) introduced cycle consistency loss to further reduce the space of possible mapping functions. This is incentivized by the idea that the learned mapping should be cycle-consistent, which is trained by the forward and backward cycle consistency losses:

$$L_{cyc}(G,F) = \mathbb{E}_{x\sim Pdata(x)}(\|F(G(x) - x\|_1) + \mathbb{E}_{y\sim Pdata(y)}(\|G(F(y) - y\|_1). \quad (4)$$

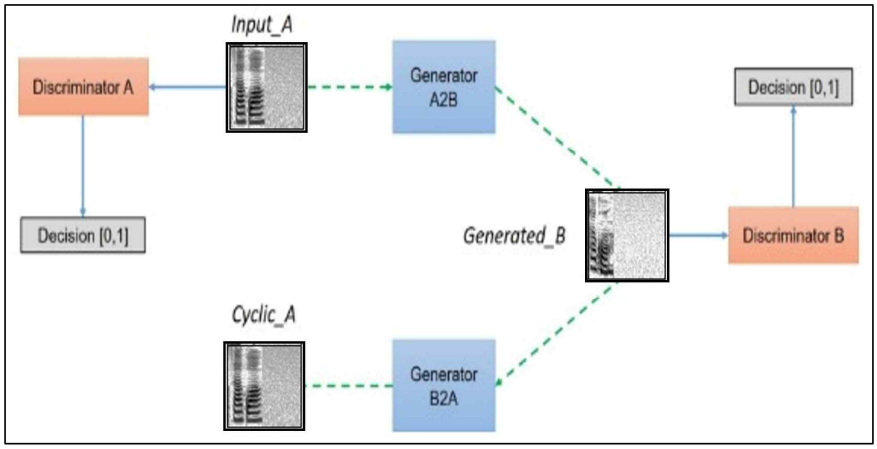


Figure 10. CycleGAN architecture which includes two generators and two discriminator neural networks. Mapping functions G : A→B and F : B→A are associated with discriminators. Discriminator B encourages G to translate A into outputs indistinguishable from domain B and vice versa for F.

Here, the network contains two mapping functions G : X→Y and F : Y→X. For each image x from domain X, the translation cycle should be able to bring x back to the original image, and vice versa. While the adversarial loss trains to match the distribution of generated images to the data distribution in the target domain, the cycle consistency losses can prevent the learned mappings G and F from contradicting each other. In the experiment in the next section, we explore the generator's behavior when trained with the conditional loss and cycle consistency loss.

4.4. Experiment

In order to test the idea that the task of translating a representation of a speech into another can be solved as a style transfer task, we compare the performances of the three algorithms; PSOLA, conditional GAN, and Cycle GAN on their abilities in translating non-native speech into native speech while maintaining the learners' voice identity. PSOLA, the traditional method in voice conversion will be the baseline. Conditional GAN is adopted as it is known for its ability in various style transfer tasks, and it would be interesting to apply it in the spectral domain. Moreover, we hypothesize that cycle consistency loss will be effective in preserving the global structure of the input spectrograms. The following section describes their implementations and the corpus.

4.4.1. Corpus

The proposed model is trained on L2KSC (L2 as Korean Speech Corpus) (Lee, 2005). The corpus is used because it is a native and non-native speech database available to the public and fits the experiment settings. We experiment with 27 hours of speech, consisting of 217 non-native speakers with 27 mother tongue backgrounds, and 107 native speakers of 54 females and 53 males. Each speaker read 300 short utterances, which are in average one second in length. When each spectrogram of non-native recording is paired with all native recordings of the same utterance, there are 1,357,321 pairs of samples for the conditional GAN training. For cycle-consistent adversarial training, there are 32,100 and 65,100 spectrograms in the native and non-native domains, each respectively. The 162 spectrograms for test are completely held-out.

4.4.2. Baseline Implementation

Baseline speech samples were generated using PSOLA algorithm, implemented in Praat (Boersma, 2001). The acoustic parameters of pitch, intensity, and duration of the native speech of the same utterance are extracted and transplanted on to the held-out non-native recordings. Its detailed algorithm is formulated in the previous Chapter.

4.4.3. Adversarial Training Implementation

For cGAN, we adopt the network architecture from Isola (2017). Its generator is an encoder-decoder network (Hinton, 2006). The input is passed through a series of layers that progressively downsample, until a bottleneck layer, at which point the process is reversed. Since there is a great deal of low-level information shared between the input and output images in style transfer, such as the location of prominent edges, it would be desirable to shuttle this information directly across the net, rather than requiring all information flow pass through all the layers, including the bottleneck. To give the generator a means to circumvent the bottleneck for information like this, skip connections were added.

For the discriminator, 70 × 70 PatchGANs are used which aim to classify whether overlapping image patches are real or fake. Since the training losses accurately capture the low frequencies (Larsen, 2015), PatchGAN is designed to restrict the discriminator to only model high-frequency structure. For

modelling high-frequencies, it is sufficient to restrict our attention to the structure in local image patches, and PatchGAN only penalizes structure at the scale of patches. Furthermore, to avoid model oscillation, in which the generator progress from one kind of sample to generating another kind of sample without eventually reaching an equilibrium, the discriminators are updated using a history of generated images rather than the ones produced by the latest generators, following the strategy in Shrivastava (2017). Adam optimizer is used (Kingma, 2015) with a learning rate of 0.0002 with a linear decay to zero after 100 epochs.

For CycleGAN, we adopt the network architecture from Zhu (2017). Its generator is an encoder-decoder network with two stride-2 convolutions, several residual blocks (He, 2016), and two fractionally-strided convolutions. Six blocks are used for 128 × 128 images and nine blocks are used for 256 × 256 higher resolution training images. The architecture is adopted from (Johnson, 2016), which has shown impressive results for style transfer and superresolution.

4.4.4. Spectrogram-to-Spectrogram Training

Trainings of the two networks are proceed in five steps: 1) native (N) and non-native (NN) speech preparation, 2) speech-to-spectrogram conversion, 3) spectrogram-to-spectrogram training, 4) inversion back into audio signal, and 5) playback of the generated audio. During the second step, audio signals were converted to spectrograms using Short-Time Fourier Transform (STFT) with windows of 512 frames and 33% overlap, which were converted to dB amplitude scale, represented using mel scale, and padded with white noise to

generate 128x128 pixels images.

Python implementation of the Griffin and Lim algorithm was used to convert the spectrogram to audio signal by using the magnitude of its STFT. It performs low-pass filtering of the spectrogram by zeroing all frequency bins above the preset cutoff frequency, and then uses the Griffin and Lim algorithm to reconstruct an audio signal from the spectrogram.

GAN is used in the third step and the conversion techniques are used during the second and the fourth steps. In order to train using GAN, the prepared samples are fed into the generator, where adversarial training is done using the discriminator which classifies whether the samples are fake (generated speech) or real (native speech). The process is shown in Figure 11. For the cycle-consistent adversarial training, there is no concatenation step, since it takes unpaired input.

During the inverse process, which is the fourth step, the Griffin Lim algorithm works to rebuild the signal with STFT such that the magnitude part is as close as possible to the spectrogram. For high quality output and minimum loss in transformations, it is run for 1,000 iterations.

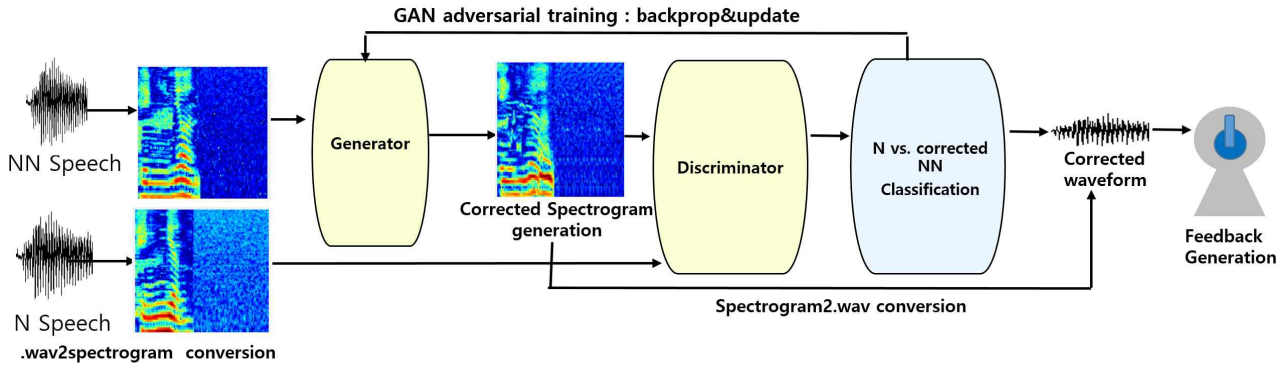


Figure 11. Framework of the proposed model: Native and non-native speech are first converted into spectrograms, which are both fed into the generator that outputs fake samples. Discriminator classifies whether the input comes from the generator or the native samples. After training, the generator model is applied to the test spectrograms, and its results are converted back into waveforms, which are played as corrective feedback to the learners.

4.5. Results and Evaluation

4.5.1. Spectrogram Generation Results

Figure 12 shows the spectrograms for non-native, generated, and native speeches at epoch 1 and epoch 3 in the conditional GAN framework. It shows that the generator quickly learns to imitate the native spectrogram by generating a fake version of the reference. After more training, the generator has discovered to generate spectrograms with higher proximity to the native. Since the test data was completely held out, this means that the model learned to recognize which word the spectrogram represents, and identified which native spectrogram should be mapped to the given non-native.

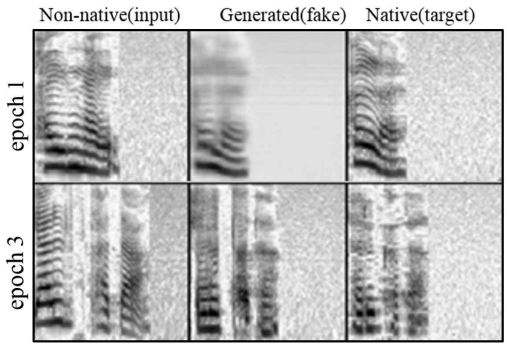


Figure 12. Comparison among input, output, and target spectrograms at epochs 1 and 3 using Pix2Pix framework.

4.5.2. Perceptual Evaluation

Since the ultimate goal is to produce fluent speech that are corrective, self-imitative, clean, and intelligible, we measure the ability of human annotators to label the generated audio. Using our three models, we generate speech samples on the test set, which amount to 486 waveforms in total. The four native Korean human raters with knowledge in linguistics were asked to listen to the original non-native utterance, followed by a generated output from one of the three models. They assigned subjective values from 1 to 5 for the following five criteria. The score of 3 was assigned if there is no difference before and after the manipulation, and 1 or 2 if the feedback resulted in a wrong correction, and 4 or 5 if the feedback was corrective.

- Holistic impression of correction: Does the generated speech correct the overall impression of the accented speech into a standard Korean accent?
- Degree of segmental correction: Does the generated speech correct the accented speech into a standard pronunciation?
- Degree of suprasegmental correction: Does the generated speech correct the accented speech into a standard intonation and prosody?
- Sound quality: Does the generated speech contain any background noise or artifacts?
- Speaker voice imitability: How similar is the generated voice to the speaker's voice?

Table 12: MOS values of perceptual test by four human experts on self-imitation feedback generation (SQ: Sound Quality)

Model	Corrective Ability			Imitability	SQ	Avg.
	Holistic	Segmental	Supra-segmental			
PSOLA	3.118	3.029	3.324	4.029	2.794	3.259
cGAN	1.970	2.485	2.152	2.697	1.636	2.188
CycleGAN	4.000	4.333	4.364	3.515	2.667	3.776

We report MOS (mean opinion scores) values in Table 12. It shows that our newly proposed CycleGAN-based speech correction method is able to generate corrective feedback. By the average score, a relative improvement of 16.67% is observed compared to the baseline PSOLA transformation. Linguistic analysis shows that the generator's corrective ability is effective both in the segmental and suprasegmental aspects. Since an error in a feedback setting can be critical in learning applications, we verified that all corrective ability scores in CycleGAN are 3 or above, which means that there was no degradation.

In addition to MOS scores, we conducted auditory transcriptions of the generated utterances on a random subsample of the test set in order to qualitatively analyze where the correction occurs. Successful cases include corrections of detensifying errors of /s͈/ in the word "fishing (낚시)." Moreover, the final rise prosodic error of the statement "It is fast (빨라요)" was corrected by the generator. Also, correcting the silence insertions between syllables, the overall rate of speech tends to be closer to the native.

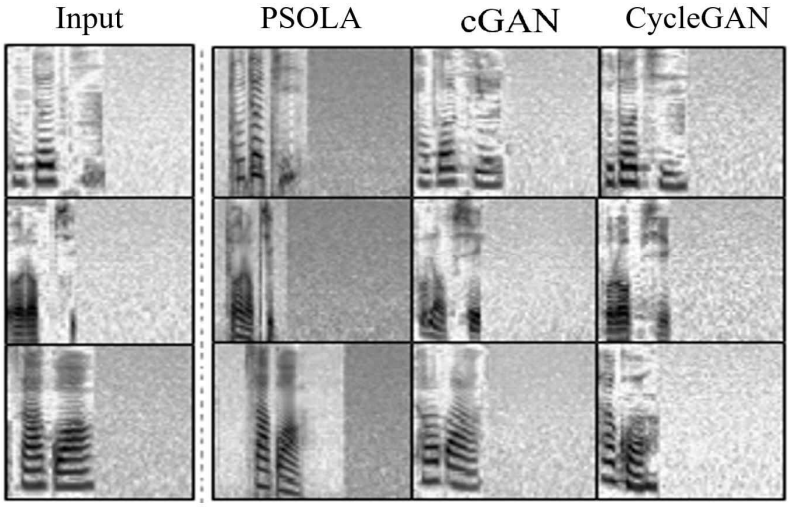


Figure 13. Input and generated spectrograms (the baseline and two GAN-based methods). Each row is a different representation of an utterance.

For the baseline PSOLA method, the evaluators report that there were numerous cases when the generated results do not make corrections, or make corrections that are perceptually trivial. On the other hand, the generated results using conditional GAN often fails to make a correction. The repositories https://github.com/sy2358/accent_conversion_GAN and https://www.youtube.com/watch?v=vYGOVabV_Y4 enable direct comparisons with auditory data.

4.5.3. Discussions

At the beginning of this Chapter, we hypothesized that accent conversion can be interpreted as a style transfer problem, which can be successfully

solved with a GAN algorithm. We also hypothesized that training the network with cycle consistency loss will induce the generator to learn a high level representation of the same underlying speech content, thanks to its forward and backward consistencies. The results confirm the hypotheses; not only the intonation, duration, and rate, but also the formant information in the spectrogram has been learnt and transplanted into the learner's voice. And since the generated speech with cycle consistency loss does not degrade the input, it can be interpreted that the loss serves to preserve the global structure of the learner's spectrogram.

The results on self-imitability show that it is able to preserve the underlying voice information and project away other accent-style information, as intended. We can ask the following question: does CycleGAN learn to preserve or project away certain information in the spectrograms? For example, can it extract the native characteristics from a speech produced by a male voice and transplant it to a speech produced by a female Chinese learner while keeping the learner's voice identity? There are possible exaplanations how CycleGAN achieves this.

One explanation is that since CycleGAN is able to learn the style of a domain, it is able to separate voice quality from the linguistic gestures. An utterance may be understood as the combination of a voice quality carrier with linguistic gestures. Being able to separate means that the network is able to deconvolve an utterance into its voice quality carrier and its linguistic filters and distinguish between-speaker variations in the dimensions of speech, such as those that are determined by physical factors, e.g. larynx size and vocal tract length. In this way, a foreign accent may be removed from an utterance by extracting its voice quality carrier and convolving it with the linguistic

gestures of a native-accented counterpart.

4.6. Summary

Automatic speech conversion/transplantation method in related works extract pitch, duration, and spectral features of the native speech and transplant them onto the learners' speech. The method allows self-imitation learning when implemented in a CAPT system, in which the learner improves the pronunciation by listening to his/her own voice. Since the method is proven to be pedagogically efficient, PSOLA transformation has been widely used for automatic corrective feedback systems.

However, most PSOLA-based feedback systems in CAPT rely on conversions at the suprasegmental level, which only extracts the duration and pitch information. This is problematic because the proficiency in a second language is fully attained only if the students have learned to modulate both the prosodic and segmental parameters equivalent to those of the native speakers. The segmental accuracy plays an important role in spoken language communication especially in L2 Korean, which was the conclusion in Chapter 3.

In this Chapter, a new methodology using a GAN that corrects both segmental and supra-segmental deviations is proposed in order to overcome this limitation. To synthesize an audio signal from the predicted spectrogram, the Griffin Lim algorithm was used to estimate a phase consistent with the predicted magnitude, followed by an inverse STFT. The perceptual evaluation shows that cycle-consistent adversarial training is a promising approach for speech correction task.

Chapter 5

Integration of Linguistic Knowledge in an Auxiliary Classifier CycleGAN for Feedback Generation

The previous chapter introduced a new method to generate corrective feedback using CycleGAN. However, although the generator seems to have acquired what and how to correct with the adversarial training, it is not necessarily the case that the information it learnt is shared with the learner. While this generator-student interaction is desirable, it would not be meaningful to merely pass the statistical distribution mapping the generator learns because the kind of information passed to the student should be linguistically motivated. The relations learnt by the generator should be linguistically representative so that the generated feedback is pedagogically meaningful.

Moreover, since it is difficult for the L2 learners to evaluate their own

pronunciations (Dlaska and Krekeler, 2008), it is helpful to provide an informative feedback. That is, it is not guaranteed that the students can perceive difference in their own speech or self-assess the pronunciation accurately. Merely listening to the speech playback in the devices with no structured linguistic content may not lead to a change in the direction closer to L2-like pronunciation. In order to make sure the feedback brings about a positive change, it is important to generate a feedback that contains linguistic information.

This chapter proposes a methodology to inject linguistic knowledge into the CycleGAN network by building dedicated generators for correction types using an auxiliary classifier. The classifier is additionally trained to distinguish three linguistic types, 'segmental' and 'suprasegmental' corrections, and 'no correction.' This forms a simple three-class convolutional neural network (CNN) (Krizhevsky, 2013), added to the feedback generation model. This Chapter therefore describes the linguistic classes and the auxiliary classifier training for the task of corrective feedback generation.

5.1. Linguistic Class Selection

The aim of this Chapter is to incorporate linguistic features into a feedback generation system. The first step in that direction is to select the linguistic feature classes. Chapter 3 in this thesis conducted an experiment with segmental, phonological, fluency, and prosody classes to find linguistically motivated features. These feature sets were designed by surveying the commonly used criteria in speech evaluation problems and considering the characteristics of Korean language.

The results from human evaluators found that all segmental and suprasegmental variations were significant predictors of speech proficiency in L2 Korean. In other words, if these human evaluators were classroom tutors, they would give feedbacks to the students both when segmental and suprasegmental error occurs. For example, if the short statement /ʥ̥adḁ/ was realized as /tɕʰadḁ/, which would be a segmental error, the learner should receive a feedback about this segmental variation. Also, if the second syllable of the word is realized with higher pitch than the first syllable, the student should receive a feedback regarding the prosody information.

Since the goal of the corrective feedback system is to mimic the human tutor as close as possible, the direction of the linguistic feedback criteria selection in this work will be to use the human evaluation results from Chapter 3. That is, both segmental and suprasegmental features will be incorporated into the feedback generation system.

5.2. Auxiliary Classifier CycleGAN Design

With the selected linguistic features, we build a classifier in order to examine whether it is possible to reliably detect the linguistic class from the generated sample. The classifier is an image classification model which can be used to classify a given spectrogram as either ‘error present’ or ‘error absent.’ In the case of the former, a linguistic class ‘segmental,’ or ‘suprasegmental’ will be assigned. The idea is to add this classifier to discriminate between the generated spectrogram of each linguistic class.

In the proposed Auxiliary Classifier CycleGAN (AC-CycleGAN), every generated sample has a corresponding class label in addition to the noise. The

auxiliary classifier gives a probability score over the class labels to discriminate between the generated samples. The cycle consistency loss induces the similarity between the generated and real spectrograms, while the classification loss induces the discriminability between linguistic classes.

The proposed AC-CycleGAN consists of three CycleGANs, each corresponding to a linguistic class, and a domain classifier (Figure 14). For each linguistic class, there is a CycleGAN with two discriminators and two mapping functions as generators, consistently with the existing CycleGAN.

The auxiliary classifier, as shown in Figure 15, is implemented as a two CNN layers with residual connections (He, 2016) and MaxPooling (Scherer et al., 2010). Rectified Linear Unit (Vinod and Hinton, 2010) is used for activation function, followed by a final linear layer classifying the samples into the three classes. Dropout of 0.5 (Hinton et al., 2014) and Adam optimizer is used. The number of epochs is initially set at 1,000 with early stopping if the model starts overfitting.

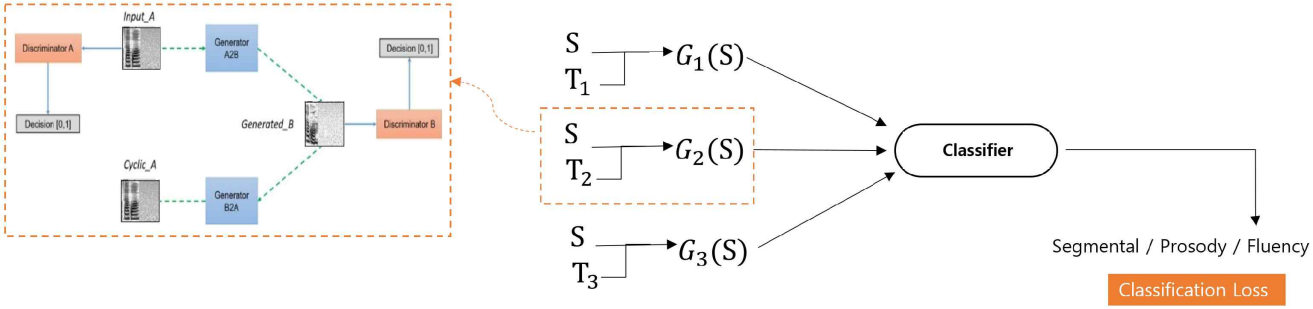


Figure 14. Proposed AC-CycleGAN architecture. The synthetic sample *Gi(S)* is generated from the source *(S)*. Cycle consistency loss is built between real samples and their corresponding reconstructed samples. The domain classifier learns to ensure the discriminability between the generated samples.

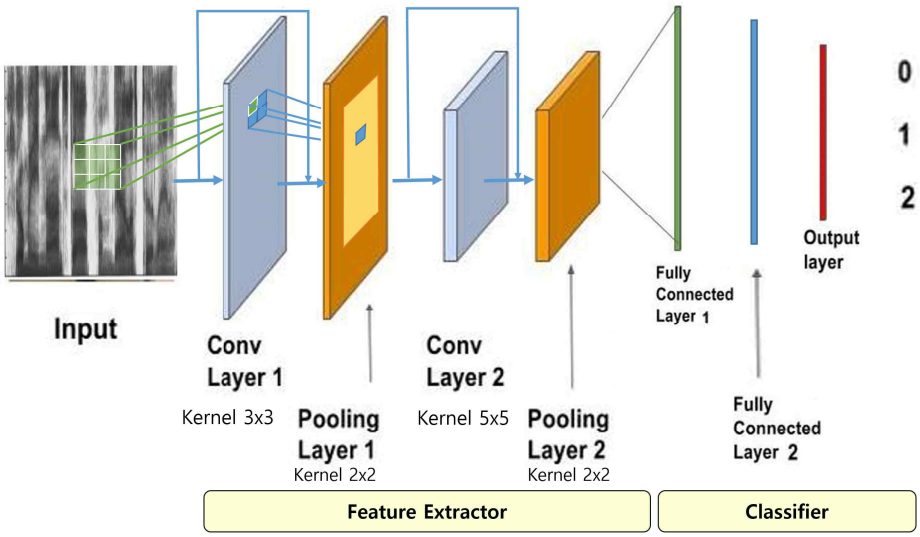


Figure 15. The auxiliary classifier architecture in the proposed AC-CycleGAN. For the 128 x 128 spectrogram sample input, the feature extractor consists of 2 convolution and pooling layers with residual connections. Then, the classifier, which consists of two fully connected and an output layer, predicts the class value between 0, 1, and 2.

5.3. Experiment and Results

5.3.1. Corpus

The proposed model is trained on L2KSC (L2 as Korean Speech Corpus) (Lee, 2005). The same corpus is used as the CycleGAN training. There are 217 non-native speakers with 27 mother tongue backgrounds, and 107 native speakers of 54 females and 53 males. Each speaker read 300 short utterances,

which are in average one second in length.

5.3.2. Feature Annotations

Unlike CycleGAN, the nature of the problem for the CNN classifier is a supervised discrimination and therefore, requires annotated data. Human annotators were asked to classify ‘error’ or ‘no error’ upon listening to the non-native speech, and in case of an error, they had to choose the error type, ‘segmental’ or ‘suprasegmental.’ Only one class was available to choose. Out of the 86.67% cases with error, 75.28% and 24.71% were classified as ‘segmental’ and ‘suprasegmental,’ each respectively.

5.3.3. Experiment Setup

The AC-CycleGAN model consists of three CycleGANs and one classifier. The architecture and the training parameters follows the description in Section 5.2. above. The CycleGAN implementation follows the description in Chapter 4. In addition, we tested the performance changes with respect to the type of loss, pretraining, fine-tuning, layer normalization (Ba et al., 2014), weight initialization, and data augmentation (Shorten et al., 2019).

First, due to the training data imbalance, weighted loss and data augmentation technique were implemented. It was mentioned that 75.28% of the errors are segmental type, and the remaining were labelled as the prosodic error. The data imbalance can be problematic, causing the model to be biased towards samples with lager distribution. A weighted loss function can be used to give more importance to the minority classes by measuring the distribution

of the data per class.

Another technique to handle the data imbalance is to augment the minority classes. In this case, the samples in the prosody class were augmented. Many kinds of augmentation are possible, such as geometric transformations, color space augmentations, mixing images, random erasing, and feature space augmentations. Given the temporal nature of the spectrogram images in this task, we only experimented the impact of spatial transformation by scaling the spectrograms horizontally by the factor of 1.1, i.e. other options like flipping or rotating the spectrograms would be less appropriate. With this scaling option, we seek to augment the data by adding more samples with higher rate of speech.

Moreover, combinations of pretraining, fine-tuning and weight initialization parameters were also implemented. Because the pretrained model from ImageNet (Deng, 2019) is not trained on any spectrogram images, but on 3.2 million images of everyday objects, animals, food and so on, it is debatable whether or not these options would contribute to the classification in this task. Enabling the pretraining and fine-tuning options allows the model to start training and fine-tune with the pretrained model. When the pretrained model is not used, the initial weights can be set to an arbitrary value of 0.5 instead of the pretrained weights.

5.3.4. Results

To evaluate the proposed method, the three class classification accuracies are reported (Table 13). The weighting and augmenting methods were tested in this experiment, as shown in the columns ‘CE Loss’ and ‘Augmentation’ in

Table 13. Auxiliary Classifier accuracies comparison of different parameter variations (CE = Cross Entropy).

CE Loss	Pretrain	Finetune	Layer -norm	Augmen-tation	Precision	Recall	F1
unweighted	No	no	no	no	0.36	0.48	0.46
weighted	No	no	no	no	0.49	0.48	0.49
weighted	Yes	yes	no	no	0.51	0.46	0.47
weighted	Yes	no	yes	yes	0.52	0.52	0.5
weighted	**Yes**	**yes**	**yes**	**yes**	**0.56**	**0.54**	**0.55**

Table 12. The results show that both methods improve the model performance, with the weighted loss being slightly more effective. For example, the confusion matrices show that before the implementations, the model tends to favor the class with more samples (Figure 16a). In contrast, the model predictions are more equally distributed after implementation of the techniques (Figure 16b).

Testing the parameters of pretraining and fine-tuning shows that the model achieves slightly better performance when the pretrained models is used both at the start of the training and fine-tuning. The best performing model is achieved when all the techniques are implemented.

Still, the model performance has a large room for improvement. Confusion patterns are found between 'No Error' and 'Prosodic' error groups, as well as 'Segmental' and 'Prosodic' error groups, suggesting that the prosodic error patterns are not easy to distinguish compared to the segmental errors. Nonetheless, the segmental errors are more easily identified. In the future work, adding more Korean native speech samples in the 'No Error' class is expected resolve the confusion patterns. Moreover, we observe that

Reference/ Predicted	No Error	Prosody	Segmental
No Error	0.00%	0.00%	21.74%
Prosody	2.17%	4.35%	39.13%
Segmental	2.17%	0.00%	30.43%

Figure. 15a

Reference/ Predicted	No Error	Prosody	Segmental
No Error	11.11%	12.35%	6.17%
Prosody	2.47%	17.28%	14.81%
Segmental	2.47%	9.88%	23.46%

Figure. 15b

Figure 16. Confusion matrices of the auxiliary classifier before and after the weighted loss and data augmentation implementations. The training data imbalance initially causes the model to be biased towards segmental errors (16a), which is alleviated and more balanced predictions are obtained (16b).

the precision, recall, and F1 in the training data reaches 0.96, 0.97, 0.96, each respectively. This confirms the ability of the classifier to learn the class distinctions, although it is more difficult to generalize to unseen data. Since the foreign language productions are necessarily influenced by the mother tongue, the performance on the unseen data can be improved by using this condition. In the future work, we plan to build a L1-dedicated classifier and test its generalizability

5.4. Summary

Although the generator seems to have acquired how to imitate the native style, it is not necessarily the case that this information is shared with the learner. Since it is difficult for the L2 learners to evaluate their own pronunciations, it is helpful to provide informative feedback on the error types. In order to enable the generator-student interaction, an auxiliary classifier is

trained to provide feedback on the linguistic error type.

Motivated by the analysis results in Chapter 3, which found that all linguistic criteria used for the human evaluation are positively correlated with human ratings, this chapter proposed an augmented variant of CycleGAN. With this simple additional 2-layer CNN, the users are expected to benefit from the knowledgeable feedback. The classifier performance is yet to be improved by accumulating more data and utilizing L1 background.

One contribution point we wish to mention before closing of this chapter is the potential of the proposed method of deep learning that preserves domain knowledge. While it is certainly efficient to let the features be learnt automatically, being able to control what it can and cannot learn, and to confirm what it has learnt is an attractive quality, especially for feedback generation tasks. Using this method that allows to work closely with linguistic analyses, and future experiments can be conducted with more fine-grained linguistic distinctions. For example, the method can be applied to a single focused error type, such as coda deletions or three-way distinctions, which are common error patterns as observed in Chapter 2. In this way, linguistic analysis results can be directly used in the automatic system, enabling individualized feedback opportunities.

Chapter 6

Conclusion

The present thesis presents a new approach for a CAPT system development, in which variation patterns and linguistic correlates with accentedness are analyzed and combined with a deep neural network approach, so that feature engineering efforts are minimized while maintaining the linguistically important factors for a corrective feedback generation task. Learning hierarchy is established by analyzing Chinese speakers' variation patterns in contrast with those of native speakers and accentedness judgement in read speech in Korean. The established priority is then modeled in an augmented Cycle-consistent generative adversarial framework.

6.1. Thesis Results

In the first part of the thesis, the pronunciation variation patterns of Korean produced by Mandarin Chinese learners were analyzed at segmental and phonological levels. Detensification and coda deletion are the most frequent phonetic characteristics, with 36.41% and 58.08% variation rates, while nasal coda insertion and non-realization of palatalization are the most frequent phonological variations.

Certain types of error deserve more importance than others and it is necessary to identify the error types that entail more perceptual value than others. This study designed a method to evaluate linguistic factors affecting L2 Korean. According to the results in the correlational analysis, segmental accuracy demonstrates the highest correlation with accentedness, followed by fluency. The takeaway of these analyses is the learning hierarchy in L2 speech Korean; coda deletions and non-realization of stress in the three-way distinctions deserve priorities in corrective feedback design of the CAPT system, followed by prosodic errors.

In the second part of the thesis, a new deep generative method for an automatic self-imitating speech correction system was proposed. The perceptual evaluation comparing PSOLA, cGAN, and CycleGAN performances shows that cycle-consistent adversarial training is a promising approach for speech correction task, outperforming the traditional method by a relative improvement of 16.67%. Then, the CycleGAN model was augmented by adding a linguistic auxiliary classifier. In addition to the generated corrected speech, the task of the classifier is to identify the type of error. The linguistic classes are adopted from the hierarchy and correlation analyses results obtained in the first part of the study. With this additional 2-layer CNN, the users are expected to benefit from the knowledgeable

feedback.

6.2. Thesis Contributions

The experiments presented in this thesis allow the results mentioned in the previous section, which contribute to the research in CAPT by conducting a large-scale linguistic analysis of L2 Korean and by proposing a novel method.

1. To the best of my knowledge, this is the first large scale corpus-based analysis to have studied the phonological variations in non-native Korean. A corpus-based analysis was conducted with larger number of utterances and balanced speaker levels. The results of this study were used to guide the priorities in teaching Korean speech to Chinese learners.
2. Generative adversarial training can learn to correct segmental errors, in addition to pitch and duration errors. The traditional PSOLA transformation is limited to pitch, duration, and intensity corrections, which is problematic because the proficiency in a second language is fully attained only if the students have learned to modulate both the prosodic and segmental parameters. This work proposed a new methodology to overcome this limitation by suggesting a model that corrects both segmental and supra-segmental deviations. automatic self-imitating speech correction system for pronunciation training. This is especially meaningful for L2 Korean, in which segmental accuracy plays an important role.
3. The AC-CycleGAN proposed in this study allows to work closely

with linguistic analyses and machine learning. In this way, linguistic analysis results can be directly used in the automatic system, enabling individualized feedback opportunities. Considering that a possible criticism of deep neural network learning method is the loss of domain knowledge, the method can be useful for combining domain knowledge and the state-of-the-art machine learning approaches and furthermore, letting the state-of-the-art machine learning discover what had been unknown in the domain knowledge.

6.3. Recommendations for Future Work

The present work conducted a large-scale linguistic analysis of L2 Korean and by proposed a novel method for CAPT. Yet, there are open issues within the field of linguistic analysis and speech generation.

The linguistic analysis approach was shown to be helpful in establishing the learning hierarchy in L2 Korean. In the future work, rater specificities can be further considered. Although the scores have been averaged per utterance in order to figure out the overall trends, scores varied among the raters. The results can be further analyzed independently of the rater-specific factors.

CycleGAN algorithm performed well on the non-native to native speech transformations. However, there is a room for improvement in CycleGAN's sound quality and speaker imitability scores. The former may be related to the lossy Griffin Lim inversion, and the artifacts produced during the process. A neural vocoder, such as WaveNet, which has been shown to significantly improve synthesis fidelity (Oord, 2016), can be tried in the future work.

The speaker voice imitability could be improved by implementing more

conditioning strategies. This may be due to the diversity in reference styles, and future work can be expended to better imitate speaker voice characteristics. For example, the current model had little controlling of the voice characteristics of input speech, and such situations can be avoided by introducing another auxiliary classifier and training the encoder and decoder so that the attribute classes of the decoder outputs are correctly predicted by the classifier. This in turn may also avoid producing buzzy-sounding speech by simply transplanting the spectral details of the input speech into its converted version.

Future experiments can be conducted using AC-CycleGAN with more fine-grained linguistic distinctions, exploiting its ability to connect linguistic analysis and machine learning methods. For example, the method can be applied to a single focused error type, such as coda deletions or three-way distinctions, which are common error patterns presented in this thesis.

Finally, considering that the purpose of the current work is in view of a pronunciation training application, the feasibility of a real-time interactive response generation needs to be tested, including, but not limited to parallelization techniques using GAN algorithms. By the results of the current study, which proves both segmental and suprasegmental corrective abilities, such effort seems worthy of future work.

Bibliography

J. C. Alderson, C. Clapham, and D. Wall. Language test construction and evaluation. Ernst Klett Sprachen, 1995.

M. A. R. Alif, S. Ahmed, and M. A. Hasan. Isolated Bangla handwritten character recognition with convolutional neural network. In *Proceedings of 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6. IEEE, 2017.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

A. Bearman, K. Josund, and G. Fiore. Accent conversion using artificial neural networks. Technical report, Stanford University, Tech. Rep, 2017.

J. Bernstein, A. Van Moere, and J. Cheng. Validating automated speaking tests. *Language Testing*, 27(3):355–377, 2010.

F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia. Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to

Hearing Impaired Speech and Speech Separation. In *Proceedings of Interspeech*, pages 4115–4119, 2019.

M. P. Bissiri, H. R. Pfitzinger, and H. G. Tillmann. Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis. In *Proceedings of 11th Australian International Conference on Speech Science & Technology*, pages 24–29. University of Auckland New Zealand, 2006.

B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, 1984.

P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9):341–345, 2001.

M. G. Busa and A. Stella. Methodological perspectives on second language ` prosody, 2012.

H. Chang, J. Lu, F. Yu, and A. Finkelstein. Paired cyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 40–48, 2018.

K. Chang. *A comparative study on phonology the phenomena between Korean and Chinese*. Master's thesis, Gangneung Wonju National University, Wonju, Korea, 2014.

F. Charpentier and E. Moulines. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Proceedings of 1st European Conference on Speech Communication and Technology*, 1989.

L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian. End-to-end neural network based automated speech scoring. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pages 6234– 6238. IEEE, 2018.

N. F. Chen, V. Shivakumar, M. Harikumar, B. Ma, and H. Li. Large-scale characterization of mandarin pronunciation errors made by native speakers of European languages. In *Proceedings of Interspeech*, pages 2370–2374, 2013.

N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li. Large-scale characterization of nonnative mandarin Chinese spoken by speakers of European origin: Analysis on ICALL. *Speech Communication*, 84:46–56, 2016.

H. Cho. A study on the teaching method of Korean pronunciation for foreigners focused on Chinese-speaking learners of Korean. *Korean Journal of General Education*, 7(6):531–559, 2013.

A. Chung. Foreign student numbers grow a record 19% in a year. https://www.universityworldnews.com/post.php?story=20181011124906535, *University World News*, 2018.

T. Chung. *A comparative study of the Korean and Chinese phonological system: focusing on the exploration of the pronunciation teaching program.* Master's thesis, Konyang University, Nonsan, Korea, 2014.

A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

J. Cui. Teaching Korean pronunciation for Chinese learners. *Bilingual Research*, 20:309–343, 2002.

J. Cui. An approach to teaching consonants to Chinese learners of the Korean language. *Keimyung Korean Studies*, 31:215–232, 2004.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A largescale hierarchical image database. In *Proceedings of IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

T. M. Derwing and M. J. Munro. Accent, intelligibility, and comprehensibility: Evidence from four l1s. *Studies in second language acquisition*, 19(1):1– 16, 1997.

M. Eskenazi. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844, 2009.

M. Eskenazi and S. Hansma. The fluency pronunciation trainer. In *Proceedings of STiLL Workshop*. Citeseer, 1998.

G. Fant. *Acoustic theory of speech production*. Number 2. Walter de Gruyter, 1970.

D. Felps, H. Bortfeld, and R. Gutierrez-Osuna. Foreign accent conversion in computer assisted pronunciation training. *Speech communication*, 51(10):920– 932, 2009.

J. E. Flege. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92:233–277, 1995.

H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda. Eduspeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*,

27(3):401–418, 2010.

L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

S. Goddijn and D. Binnenpoorte. Assessing manually corrected broad phonetic transcriptions in the spoken Dutch corpus. In *Proceedings of 15th ICPhS*, pages 1361–1364, 2003.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

J. Han. *Teaching Korean pronunciation*. Hollym, Seoul, Korea, 2003.

T. K. Hansen. Computer assisted pronunciation training: The four 'k's of feedback. *Current Developments in Technology-Assisted Education*, pages 342–346, 2006.

A. Haque, M. Guo, and P. Verma. Conditional end-to-end audio transforms. In *Proceedings of Interspeech*, pages 2295–2299, 2018.

A. M. Harrison, W. Y. Lau, H. M. Meng, and L. Wang. Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer. In *Proceedings of Interspeech*, pages 2787–2790, 2008.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore,

D. Higgins, X. Xi, K. Zechner, and D. Williamson. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, 25(2):282–306, 2011.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580,* 2012.

H. Hong and M. Chung. Improving phone recognition performance via phonetically-motivated units. In *Proceedings of Interspeech*, pages 1663–1666, 2009.

H. Hong, S. Kim, and M. Chung. A corpus-based analysis of Korean segments produced by Japanese learners. In *Proceedings of SLaTE*, pages 189–192, 2013.

H. Hong, H. Ryu, and M. Chung. The relationship between segmental production by Japanese learners of Korean and pronunciation evaluation. *Phonetics and Speech Sciences*, 6(4):101–108, 2014.

S. Hong. *Juncture patterns in Chinese learners of Korean*. Master's thesis,

Korea University, Seoul, Korea, 2016.

W. Hu, Y. Qian, F. K. Soong, and Y. Wang. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154–166, 2015.

M. Hwang. *A Study on Errors in Consonant Pronunciation by and an Educational Approach for Korean Language Learners from Chinese-Speaking Regions*. Ph.D. thesis, Thesis, Chungnam National University, Daejon, Korea, 2012.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of International Conference on Machine Learning*, pages 448–456, 2015.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of European conference on computer vision*, pages 694–711. Springer, 2016.

M. Jung. Strategic pronunciation education for Chinese learners of Korean, Korean language studies. *The Society of Korean Language and Literature,* 38:345– 369, 2008.

Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-backward error: Automatic detection of tracking failures. In *Proceedings of 20th International Conference on Pattern Recognition*, pages 2756–2759. IEEE, 2010.

T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of International Conference on Learning Representations*, 2018.

A. Kim. *A Study on Korean Language Learner's Speaking Ability Development Focusing on Chinese Learners Residing in Korea*. Ph.D. thesis, KyungHee University, Seoul, Korea, 2017.

K. Kim. *A Study on Teaching Korean Pronunciation to Chinese Speakers*. Ph.D. thesis, Dankook University, Chonan, Korea, 2008.

S. Kim. and H. Jung. A study on the utilization of speech recognition technology in foreign language learning applications - focusing on English and French speech. *The Digital Contents Society*, 19.4: 621-630, 2018.

Y. Kim, S. Nam, S. Lee, and S. Lee. A study on the correlation between proficiency and productive ability of learners of Korean. *The Society of Korean Language and Literature*, 164:209–244, 2013.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations*, 2015.

W. B. Kleijn and K. K. Paliwal. *Speech Coding and Synthesis.* Elsevier Science Inc., USA, 1995.

R. W. Kortekaas and A. Kohlrausch. Psychoacoustical evaluation of the pitch synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli. *The Journal of the Acoustical Society of America*, 101(4):2202–2213, 1997.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedigs Advances in neural information processing systems*, pages 1097–1105, 2012.

A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300,* 2015.

A. Lee and J. R. Glass. Context-dependent pronunciation error pattern discovery with limited annotations. In *Proceedings of 5th Annual Conference of the International Speech Communication Association*, 2014.

B. Lee. *A study on the dynamic development of complexity, accuracy and fluency in Korean learners' writing and speaking.* Ph.D. thesis, Yonsei University, Seoul, Korea, 2016.

H.-Y. Lee. *Korean phonetics*. Taehaksa, Seoul, Korea, 1996.

K.-N. Lee and M. Chung. Morpheme-based modeling of pronunciation variation for large vocabulary continuous speech recognition in Korean. *IEICE transactions on information and systems*, 90(7):1063–1072, 2007.

L. Leng. *Chinese Pronunciation of Consonants Education of Korea Students.* Ph.D. thesis, Chungnam National University, Daejeon, Korea, 2014.

X. Li, S. Mao, X. Wu, K. Li, X. Liu, and H. Meng. Unsupervised discovery of nonnative phonetic patterns in l2 English speech for mispronunciation detection and diagnosis. In *Proceedings of Interspeech*, pages 2554–2558, 2018.

Y.-H. Lin. *The Sounds of Chine*se. Cambridge University Press, USA, 2007.

M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Proceedings of Advances in neural information processing systems*, pages 700–708, 2017.

F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4990–4998, 2017.

P. Martin. Winpitch, a multimodal pronunciation software. In *Proceedings of InSTIL/ICALL Symposium*, 2004.

K. McBride. Which features of Spanish learners' pronunciation most impact listener evaluations? *Hispania*, pages 14–30, 2015.

W. Menzel, D. Herron, P. Bonaventura, and R. Morton. Automatic detection and correction of non-native English pronunciations. *Proceedings of INSTILL*, pages 49–56, 2000.

F. Metze. *Articulatory features for conversational speech recognition*. Ph.D. thesis, Verlag nicht ermittelbar, 2005.

P. F. d. V. Muller. *Automatic oral proficiency assessment of second language speakers of South African English*. Ph.D. thesis, Stellenbosch: University of Stellenbosch, 2010.

M. J. Munro and T. M. Derwing. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning,* 45(1):73–97, 1995.

K. Nagano and K. Ozawa. English speech training using voice conversion. In

Proceedings of 1st International Conference on Spoken Language Processing, pages 1169–1172, 1990.

V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of 27th international conference on machine learning,* pages 807–814, 2010.

A. Neri, C. Cucchiarini, and H. Strik. Segmental errors in Dutch as a second language: how to establish priorities for CAPT. *iCALL symposium.* 2004.

A. Neri, C. Cucchiarini, H. Strik, and L. Boves. The pedagogy-technology interface in computer assisted pronunciation training. *Computer assisted language learning,* 15(5):441–467, 2002.

M. G. O'Brien. L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning,* 64(4):715–748, 2014.

A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of 34th International Conference on Machine Learning*.V 70: 2642–2651, 2017.

Y. R. Oh, J. S. Yoon, and H. K. Kim. Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. *Speech Communication*, 49(1):59–70, 2007.

M. Peabody and S. Seneff. Towards automatic tone correction in non-native mandarin. In *Proceedings of International Symposium on Chinese Spoken Language Processing*, pages 602–613. Springer, 2006.

E. Pellegrino and D. Vigliano. Self-imitation in prosody training: a study on

Japanese learners of Italian. In *Proceedings of SLaTE*, pages 53–57, 2015.

M. Pettorino and M. Vitale. Transplanting native prosody into second language speech. In *Proceedings of Methodological Perspectives on Second Language Prosody*, pages 11–16, 2012.

M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. CNN architectures for large-scale audio classification. In *Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 131–135. IEEE, 2017.

K. Probst, Y. Ke, and M. Eskenazi. Enhancing foreign language tutors–in search of the golden speaker. *Speech Communication*, 37(3-4):161–173, 2002.

X. Qian, H. Meng, and F. Soong. Discriminatively trained acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT). In *Proceedings of Interspeech*, 2010.

Y. Qin. *A study on the plans for teaching Korean pronunciation to Chinese learners*. Ph.D. thesis, thesis, Konkuk university, Seoul, Korea, 2010.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of International Conference on Learning Representations*, 2016.

S. Rhee, J. Kim, and J. Chang. Design and construction of speech corpus for Korean as a foreign language (l2ksc). *The Journal of Chinese Language and Literature*, 33:35–53, 2005.

O. Ronneberger, P. Fischer, and T.-n. Brox. Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference*

on Medical Image Computing and Computer-Assisted Intervention, pages 234–241, 2015.

H. Ryu, K. Lee, S. Kim, and M. Chung. Improving transcription agreement of nonnative English speech corpus transcribed by non-natives. In *Proceedings of SLaTE*, pages 61–64, 2011.

K. Saito and Y. Akiyama. Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation*, 3(2):199– 217, 2017.

S. Sayamon. *The Influence of Learning Contexts on Speaking Development of Intermediate Thai Learners of Korean-Focusing on Fluency and Pronunciation.* Ph.D. thesis, KyungHee University, Seoul, Korea, 2016.

D. Scherer, A. Muller, and S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *Proceedings of International conference on artificial neural networks*, pages 92–101. Springer, 2010.

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

H. Strik and C. Cucchiarini. Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29(2-4):225–246, 1999.

Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200,* 2016.

Y. Tsubota, T. Kawahara, and M. Dantsuji. Recognition and verification of English by Japanese students for computer-assisted language learning system.

In *Proceedings of 7th International Conference on Spoken Language Processing*, pages 1205–1208, 2002.

J. van Doremalen. *Developing automatic speech recognition-enabled language learning applications: from theory to practice.* Ph.D. thesis, Radboud University, Nijmegen, Nederland, 2014.

S. M. Witt. *Use of speech recognition in computer-assisted language learning.* Ph.D. thesis. University of Cambridge, U.K., 1999.

S.-H. Yang and M. Chung. Linguistic factors affecting evaluation of l2 Korean speech proficiency. In *Proceedings of SLaTE*, pages 53–58, 2017.

S. H. Yang and M. Chung. Speech Assessment using Generative Adversarial Network. In *Proceedings of Machine Learning in Speech and Language Processing Workshop*, 2018.

S. H. Yang and M. Chung. Self-Imitating Feedback Generation Using GAN for Computer-Assisted Pronunciation Training. In *Proceedings of Interspeech*, pages 1881–1885, 2019.

S.-H. Yang, M. Na, and M. Chung. Modeling pronunciation variations for non-native speech recognition of Korean produced by Chinese learners. In *Proceedings of SLaTE*, pages 95–99, 2015.

C. Yoo. *Comparison of phonology between Korean and Chinese for Korean pronunciation education.* Master's thesis, Changwon University, Changwon, Korea, 2012.

K. Yoon. Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody. *Journal of the Modern British*

& American Language & Literature, 25(4):197–215, 2007.

K. Zechner, D. Higgins, X. Xi, and D. M. Williamson. Automatic scoring of nonnative spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895, 2009.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of IEEE international conference on computer vision*, pages 2223–2232, 2017.

Correct Pronunciation LITE, 바른 발음 LITE. https://play.google.com/store/apps/details?id=com.elsebook.handal.baleum.lite&hl=en_US.

Kmaru SPEECH, 한글 한국어 공부. http://www.kmaru.com/v1/.

Korean Pronunciation Training, 한국어 발음 교육. https://play.google.com/store/apps/details?id=air.kr.co.eduhansol.poppingKorean&hl=ko.

Learn Korean Phrases. https://play.google.com/store/apps/details?id=com.bravolang.korean&hl=en_US.

Pronunciation Pratice, 나도 아나운서:발음연습. https://apkdownloadforwindows.com/app/com.nowannouncer_free/.

Rosetta Stone. http://www.rosettastone.eu, 2013.

Talk to Me. http://www.auralog.com/en/Individuals talktome.htm, 2002.

Tell Me More. http://www.tellmemore.com, 2013.

Appendix

The appendix provides the supplementary information on the read speech corpus used in this thesis.

Appendix I.

List of Read Speech Words in L2KSC Corpus.

1. 앉다
2. 콧물
3. 여덟
4. 밖
5. 건강
6. 씨앗
7. 난로
8. 늦었다
9. 앞치마
10. 부탁했어요
11. 특히
12. 벌레
13. 거짓말
14. 썼다
15. 왕
16. 처음
17. 더럽다
18. 요리
19. 아버지
20. 근처
21. 좋겠다
22. 관광
23. 사과
24. 두더지
25. 삐약삐약
26. 규칙
27. 꽃
28. 너구리
29. 뻐꾸기
30. 기악
31. 등산로
32. 그렇지만
33. 토요일
34. 같아요
35. 가득하다
36. 신라호텔
37. 신선하다
38. 압력
39. 치약
40. 라디오
41. 궤도
42. 일년
43. 축하
44. 곤란합니다
45. 듣습니다
46. 복잡하다
47. 냇물
48. 결단력
49. 된장
50. 사람
51. 공부
52. 문화
53. 까마귀
54. 싸움
55. 입원료
56. 노래
57. 허리띠
58. 빨갛다
59. 끼리끼리
60. 숭늉
61. 안녕
62. 옆 집
63. 아프다
64. 넣다
65. 실내
66. 끼우다
67. 낚시
68. 전화
69. 나라

70. 핑계
71. 동화책
72. 한라산
73. 감다
74. 웃겼다
75. 극장
76. 빨래
77. 가지
78. 명동
79. 선로
80. 쑥
81. 안경
82. 텃밭
83. 기억
84. 끝났군요
85. 따뜻하다
86. 수탉
87. 덥다
88. 빨라요
89. 넉넉하다
90. 멀다
91. 깨끗하다
92. 신문로
93. 즐겁다
94. 학년
95. 쓰다
96. 문을 닫는다.
97. 삯
98. 고맙습니다
99. 다섯 명
100. 낯
101. 선물
102. 벗
103. 개미
104. 디딤돌
105. 돼지
106. 옷깃
107. 닫히다
108. 연락하세요
109. 달립니다
110. 오빠
111. 4000 원입니다.
112. 송아지
113. 졸업식
114. 옛날
115. 아빠
116. 없다
117. 예습
118. 빵집
119. 혓바닥
120. 회의
121. 읽기
122. 김밥
123. 냉큼
124. 멀리
125. 횃불
126. 한 냥
127. 꾀꼬리
128. 멋있다
129. 하숙집
130. 경복궁
131. 꺼칠하다
132. 뚜벅뚜벅
133. 꽈리고추
134. 밤낮
135. 21 일입니다
136. 낱개
137. 읊다
138. 조약돌
139. 떼쓰다
140. 한국말
141. 책
142. 미국
143. 연필
144. 꾸지람
145. 누나
146. 통화
147. 동생

148. 닮았지요
149. 먹다
150. 무엇입니까
151. 부엌
152. 저녁식사
153. 10 월
154. 젊다
155. 압구정동
156. 불렀어요
157. 나뭇잎
158. 힘줄
159. 넓다
160. 밥
161. 민주주의의 의의
162. 들어가세요
163. 있어요
164. 생각했습니다
165. 크다
166. 파랑
167. 얌냠
168. 폐렴
169. 어머니
170. 주었다
171. 푼돈
172. 길러요
173. 꽃집
174. 자장면
175. 거리
176. 뿌리
177. 돈을 찾아요
178. 단어
179. 귀
180. 칼날
181. 낮잠
182. 편리합니다
183. 놀라다
184. 수영하다
185. 북녘
186. 값도
187. 갑자기
188. 어른
189. 짬뽕
190. 게시판
191. 장독
192. 삿갓
193. 목욕
194. 승리
195. 괜찮습니다
196. 냉면
197. 벼 이삭
198. 젓가락
199. 옷
200. 저녁
201. 괘종시계
202. 달라요
203. 고장
204. 티끌
205. 부르다
206. 달걀
207. 무쇠
208. 그림
209. 꽤 많다
210. 울다
211. 핥다
212. 촛불
213. 느티나무
214. 선생님
215. 골라요
216. 계란
217. 전라도
218. 교회
219. 우유
220. 몫
221. 늦잠
222. 싫군요
223. 꿩
224. 신라면
225. 초록색

226. 맵시
227. 숙녀
228. 삶다
229. 춥다
230. 국화
231. 독립
232. 팔다
233. 캄캄하다
234. 사랑하다
235. 좌회전
236. 껴안다
237. 월요일
238. 두뇌
239. 공항
240. 연습하기
241. 봄
242. 갔다
243. 얘기
244. 예쁘다
245. 국제
246. 밟다
247. 가족
248. 갸륵하다
249. 한강
250. 할 일
251. 키읔
252. 과자
253. 좋습니다
254. 코 골아요
255. 밝아요
256. 권리
257. 결혼
258. 못했어요
259. 땅콩
260. 효도
261. 괴물
262. 달나라
263. 흐린 날
264. 백 년
265. 빼앗다
266. 세상
267. 맛있다
268. 몇 년
269. 물약
270. 모양
271. 국물
272. 음악
273. 번호
274. 토끼
275. 잉어
276. 신발
277. 생산량
278. 함께
279. 선릉역
280. 해돋이
281. 길어요
282. 컴퓨터
283. 겨울
284. 튀김
285. 맞는다
286. 답답하다
287. 애기
288. 꿰매다
289. 휴게실
290. 옮다
291. 아홉시
292. 당뇨병
293. 냇가
294. 줄넘기
295. 돌솥 비빔밥
296. 십 년
297. 입학
298. 숨다
299. 이쑤시개
300. 우산

Appendix II.

List of Read Speech Sentences used for Accentedness Rating Task in Chapter 3 of the thesis.

1. 어머니께서 어디에 계십니까?
2. 제 취미는 책 읽기입니다.
3. 지금은 공부하지 않고 쉬고 싶어요.
4. 중국 음식 중에서 뭐가 유명해요?
5. 극장에 들어갈 때 음악이 나왔어요.
6. 값이 싸고 질이 좋은 물건을 팔아요.
7. 다 준비했으니까 걱정할 필요없어요.
8. 요즘은 포장이사를 하는 사람이 많습니다.
9. 우리 선생님은 벌써 결혼했습니다.
10. 보통 집에 오자마자 밥부터 먹어요.
11. 한국에 몇 번 와 봤어요?
12. 혼자 외국 여행을 한 적이 있어요?
13. 돈이 없는데 좀 빌려 주세요.
14. 식사를 주문한 후에 음료수를 시켰어요.
15. 찬바람이 불어서 감기에 걸렸어요.
16. 먼저 앞에 있는 닭고기를 볶으세요.
17. 같이 있으니까 기분이 더 좋네요.
18. 고기를 먹지 않는 사람도 많다.
19. 어머니는 아들을 낳고 기뻐했습니다.
20. 모르는 사람이 많아서 어색해요.
21. 값도 싸고 질도 좋아요.

22. 머리를 짧게 자르고 치마를 입으니까 다른 사람 같아요.
23. 깨끗한 사람인 줄 알았는데 그렇지 않더라.
24. 좋아하는 사람한테서 꽃 한 송이 받았으면 좋겠어요.
25. 내일은 늦게 와도 괜찮아요.
26. 게으른 학생은 우수한 성적을 받을 수 없지요.
27. 신부는 꽃다발을 들고 있고 신랑은 꽃을 꽂고 있어요.
28. 며칠 동안 청소를 못해서 방이 굉장히 더럽다.
29. 식사 예절에 대해서 발표해 보도록 합시다.
30. 나무가 너무 커서 자르는 데 한 시간이나 걸렸다.
31. 실례지만 시청에 가려면 몇 번 버스를 타야 돼요?
32. 꽃무늬 원피스가 예뻐 보여서 한 벌 샀어요.
33. 친구는 빨간색 옷을 입으면 진짜 잘 어울려요.
34. 형제 중에서 맏이가 제일 힘든 것 같아요.
35. 정희 씨는 전공이 뭐예요?
36. 주말에는 백화점에서 쇼핑을 하거나 집에서 책을 읽습니다.
37. 여학생들에게 심리학이나 법학이 인기가 있다는군요.
38. 제 가방에 책, 공책, 연필이 있습니다.
39. 냉장고 안에 과일과 채소가 많지만 먹고 싶지 않아요.
40. 우리 하숙집은 신촌에 있고 옆에는 공원이 있어요.
41. 아파트 근처에 병원하고 약국하고 편의점이 있습니다.
42. 값 비싼 명품도 좋지만 오래 기억할 수 있는 선물이 좋지요.
43. 대학 동창들과 지금도 연락하면서 안부를 주고 받는다.
44. 언니는 노래도 잘 부르고 피아노도 잘 쳐요.
45. 서로 생각은 다르지만 맞추려고 노력한다.
46. 한식집에는 냉면, 김치찌개, 비빔밥, 갈비탕이 있습니다.

47. 아직 우리는 건강하고 젊기 때문에 더 많은 것들을 할 수 있지요.
48. 동생 생일에 케이크도 사고 축하 노래도 불렀어요.
49. 설날과 추석은 한국의 가장 큰 명절입니다.
50. 3 월이 되니까 햇볕이 참 따뜻합니다.
51. 일요일마다 봉사활동을 하면서 큰 보람을 느꼈다.
52. 아이들을 가르치기가 힘들지만 재미있고 보람도 있답니다.
53. 저는 날마다 아침 일곱 시부터 여덟 시까지 중국어를 배워요.
54. 오늘 오후에 회사에서 친구와 약속이 있어요.
55. 수업이 끝나고 좀 늦게 식료품 가게에 갔습니다.
56. 오늘은 거래처에 갔다가 특별한 손님을 만났습니다.
57. 횡단보도를 건너서 학교 쪽으로 올라가면 정류장이 있어요.
58. 먼저 깨소금, 고춧가루 등 기본적인 양념을 준비하세요.
59. 큰 오빠는 무역회사에서 일하고 작은 오빠는 신문사에서 근무합니다.
60. 음식물 쓰레기는 반드시 분리해서 배출해야 돼요.
61. 금연구역에서 담배를 피우면 벌금을 내게 합시다.
62. 뭐니뭐니 해도 업무 능력이 특히 중요하지요.
63. 공항은 출국하는 사람과 입국하는 사람들로 북적거렸다.
64. 시가행진으로 차가 밀려서 약속 시간에 늦었어요.
65. 날씨가 나빠서 공연 계획을 취소할 수밖에 없었습니다.
66. 동대문 시장이나 남대문 시장은 외국인들이 많이 찾는 곳입니다.
67. 값을 치르기 전에 유통기한을 확인하세요.
68. 국립학교 등록금이 사립학교 등록금에 비해서 싼 편이에요.
69. 도시 출신과 시골 출신 직원의 장점을 살릴 계획이다.
70. 때이른 더운 날씨로 해수욕장은 개장을 서두르고 있다.

71. 우리 팀 전력이 상대편 전력보다 훨씬 뒤떨어집니다.
72. 옛날에 목욕탕이었던 곳들이 대부분 찜질방으로 바뀌었어요.
73. 대학 1 학년생은 1997 년생이 제일 많다고 합니다.
74. 건물 앞은 복잡하니까 지하철 역 입구에서 내립시다.
75. 김치를 담글 줄 아는 사람이 줄어들고 있어요.
76. 필요한 도움을 드리지 못해서 안타깝습니다.
77. 1 년 동안 같이 공부한 친구들과 이별하기가 아쉽다.
78. 치킨과 맥주를 함께 먹는 것이 유행이라고 하네요.
79. 노력한 만큼 좋은 결과가 있을까요?
80. 대사관이나 영사관에 연락하면 중요한 정보를 얻을 수 있습니다.
81. 물론 전기 밥솥이 압력 밥솥에 비해서 편리하지요.
82. 건강을 잃으면 건강을 되찾기까지 오랜 시간이 걸려요.
83. 아침 서울은 바람도 불고 구름도 잔뜩 꼈습니다.
84. 습기가 많아서 그런지 조금만 움직여도 땀이 비 오듯 흘러요.
85. 올림픽에 참가한 국가의 선수들은 특별한 보호를 받습니다.
86. 죽을 만큼 힘들지만 포기하지 않을 거예요.
87. 모두가 예측한 대로 가까운 친척이 범인으로 밝혀졌어요.
88. 세탁기가 없으니까 빨래가 많이 밀려서 힘들어요.
89. 정치적인 책임은 대통령에게 있다고 봅니다.
90. 처음에는 정말 많이 넘어졌는데 이제는 스키 타기가 어렵지 않습니다.
91. 논문 주제와 관련된 자료를 찾으려고 인터넷을 검색했어요.
92. 7 월 중순이면 장마가 끝나고 무더위가 시작된다고 합니다.
93. 태풍 피해자가 정부를 상대로 피해 보상을 요구했습니다.
94. 관리사무소는 몰리는 주민들로 골머리는 앓는다고 합니다.

95. 해결할 수 없는 사례만 늘어나고 있어서 국민들이 두려워하고 있다.
96. 퇴직 후의 생활을 염려하는 중장년층이 해마다 늘고 있어요.
97. 운전면허증을 받기 위해서 운전학원에 등록했어요.
98. 아이들은 장래 희망으로 연예인을 1 순위로 꼽았어요.
99. 여러분은 어떻게 만든 음식을 드시고 계십니까?
100. 가장 많이 팔리는 음료수는 탄산음료로 나타났습니다.

초록

외국어로서의 한국어 교육에 대한 관심이 고조되어 한국어 학습자의 수가 크게 증가하고 있으며, 음성언어처리 기술을 적용한 컴퓨터 기반 발음 교육(Computer-Assisted Pronunciation Training; CAPT) 어플리케이션에 대한 연구 또한 적극적으로 이루어지고 있다. 그럼에도 불구하고 현존하는 한국어 말하기 교육 시스템은 외국인의 한국어에 대한 언어학적 특징을 충분히 활용하지 않고 있으며, 최신 언어처리 기술 또한 적용되지 않고 있는 실정이다. 가능한 원인으로써는 외국인 발화 한국어 현상에 대한 분석이 충분하게 이루어지지 않았다는 점, 그리고 관련 연구가 있어도 이를 자동화된 시스템에 반영하기에는 고도화된 연구가 필요하다는 점이 있다. 뿐만 아니라 CAPT 기술 전반적으로는 신호처리, 운율 분석, 자연어처리 기법과 같은 특징 추출에 의존하고 있어서 적합한 특징을 찾고 이를 정확하게 추출하는 데에 많은 시간과 노력이 필요한 실정이다. 이는 최신 딥러닝 기반 언어처리 기술을 활용함으로써 이 과정 또한 발전의 여지가 많다는 바를 시사한다.

따라서 본 연구는 먼저 CAPT 시스템 개발에 있어 발음 변이 양상과 언어학적 상관관계를 분석하였다. 외국인 화자들의 낭독체 변이 양상과 한국어 원어민 화자들의 낭독체 변이 양상을 대조하고 주요한 변이를 확인한 후, 상관관계 분석을 통하여 의사소통에 영향을 미치는 중요도를 파악하였다. 그 결과, 종성 삭제와 3중

대립의 혼동, 초분절 관련 오류가 발생할 경우 피드백 생성에 우선적으로 반영하는 것이 필요하다는 것이 확인되었다.

교정된 피드백을 자동으로 생성하는 것은 CAPT 시스템의 중요한 과제 중 하나이다. 본 연구는 이 과제가 발화의 스타일 변화의 문제로 해석이 가능하다고 보았으며, 생성적 적대 신경망 (Cycle-consistent Generative Adversarial Network; CycleGAN) 구조에서 모델링하는 것을 제안하였다. GAN 네트워크의 생성모델은 비원어민 발화의 분포와 원어민 발화 분포의 매핑을 학습하며, Cycle consistency 손실함수를 사용함으로써 발화간 전반적인 구조를 유지함과 동시에 과도한 교정을 방지하였다. 별도의 특징 추출 과정이 없이 필요한 특징들이 CycleGAN 프레임워크에서 무감독 방법으로 스스로 학습되는 방법으로, 언어 확장이 용이한 방법이다.

언어학적 분석에서 드러난 주요한 변이들 간의 우선순위는 Auxiliary Classifier CycleGAN 구조에서 모델링하는 것을 제안하였다. 이 방법은 기존의 CycleGAN에 지식을 접목시켜 피드백 음성을 생성함과 동시에 해당 피드백이 어떤 유형의 오류인지 분류하는 문제를 수행한다. 이는 도메인 지식이 교정 피드백 생성 단계까지 유지되고 통제가 가능하다는 장점이 있다는 데에 그 의의가 있다.

본 연구에서 제안한 방법을 평가하기 위해서 27개의 모국어를 갖는 217명의 유의미 어휘 발화 65,100개로 피드백 자동 생성 모델을 훈련하고, 개선 여부 및 정도에 대한 지각 평가를 수행하였다. 제안된 방법을 사용하였을 때 학습자 본인의 목소리를 유지한 채 교정된 발음으로 변환하는 것이 가능하며, 전통적인

방법인 음높이 동기식 중첩가산 (Pitch-Synchronous Overlap-and-Add) 알고리즘을 사용하는 방법에 비해 상대 개선률 16.67%이 확인되었다.

Acknowledgements

First, my thanks go to my advisor, Professor Minhwa Chung, for giving me the opportunity to research in the Spoken Language Lab. I could not have wished for an advisor who could be so supportive throughout my time in SNU. While providing me with incredibly important opportunities, he introduced me to the combination of science and language. I thank him for letting me pursue my ideas, while keeping me on the right path. His clear thinking and willingness to give me the time and support I needed during one of the busiest years of his life allowed me go through to the finish.

I want to thank Kyuwhan Lee for sharing his insights with me. Although an individual work, his influence was essential for this thesis. I regret that I can no longer share my research ideas and results with him, and that he will not be able to see me graduate – nevertheless, his memory will be eternal.

This thesis profited greatly from discussions with my thesis committee, Professor Hong-Gee Kim, Professor Ho-Young Lee, Professor Sunhee Kim, and Professor Ji-Hwan Kim. Invaluable thanks goes to Dr. Jean Senellart for discussing and growing our ideas, and sharing with me something much more than the academic passion over the course of time.

A part of the thesis has been conducted towards the course of research project with ETRI. I was never bored a minute while working on this research project; be they working on the pronunciation evaluation and training system, or enjoying the workshops held in beautiful places in Korea. I especially

thank Professor Yongjoo Lee for his unwavering support, and also Professor Seok-Chae Lee, Dr. Jeonkyu Park, Dr. Hyungbae Jeon, Dr. Yoori Oh, and Dr. Yoongkyung Lee. I learnt a great deal about engineering and machine learning from the lectures of Professor Chung, Professor Seongwoo Kim, Professor Zhang, and Professor Seong, thanks to the technical practices and programming projects.

This thesis would not have been possible without the academic and emotional support of Taehyeong Kim. More thanks are also due to Hyunwoong Ko and Sungjae Cho, for making the school an exciting, challenging, and fun place to be. I hope to continue with our collaboration to achieve something greater together.

I owe thanks to all members of my lab, including Dr. Minsu Na, Jong In, Jooyoung, Seoha, Jeemin, Eun Jung, Abner, Seogyeong, Mikyoung, and Seunghun, for their superb maintenance and for being such a wonderful company every day. Thanks also go to my friends and colleagues; Seong Hee, Hajin, Haerim, Jeemin, Geunyoung, Jeehyun, Jeeun, Gicheon, and Petar. I will never forget the dozens of interesting conversations, our joys and cheers.

Finally, this thesis is dedicated to my parents, Jaichang Yang and Yeonsun Wang, for their endless kindness, patience, and love. They reminded me that there is much more in life to appreciate than doing a better machine learning. To my grandmother, Junghye Ji, my sister, Jisun Yang, my brother-in-law, Taejin Park, and my niece, Hyunjin Park: thank you for giving me the motivation, persistence, and smiles.

To all of the above and to anyone I unintentionally missed who has touched my life and brought me to this point: Thank you.